E-ISSN : 2614-8404
P-ISSN : 2776-3234

JISA

(JURNAL INFORMATIKA dan SAINS)

VOL 4 No 2
December

Published by :
Program Studi Teknik Informatika
UNIVERSITAS TRILOGI

2021

JISA
(Jurnal Informatika dan Sains)

Volume 4, Edition 2, December 2021

Published by:
Program Studi Teknik Informatika
Universitas Trilogi

JISA	Vol : 4	Ed :2	Page : 96-182	Jakarta, Jun 2021	e-ISSN: 2614-8404	p-ISSN: 2776-3234

JISA
(Jurnal Informatika dan Sains)

Volume 4, Edition 2, December 2021

Editorial Address
Ruang Dosen Fakultas Industri Kreatif dan Telematika Lantai 3
Jalan Taman Makam Pahlawan No. 1, Kalibata, Pancoran, RT.4/RW.4, Duren Tiga, Pancoran, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12760Telp :(021) 798001

Published by:
Program Studi Teknik Informatika
Universitas Trilogi

JISA	Vol : 4	Ed :2	Page : 96-182	Jakarta, Des 2021	e-ISSN: 2614-8404	p-ISSN: 2776-3234

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Table of Content

e-ISSN: 2614-8404
p-ISSN :2776-3234

Climate Prediction Using RNN LSTM to Estimate Agricultural Products Based on Koppen Classification

Novia Andini[1*)], Wiranto Herry Utomo[2]
Information Technology, Faculty of Computing, President University
Email: [1]novia.andini@student.president.ac.id, [2]wiranto.herry@president.ac.id

Abstract *–* The yield of an agricultural process is very important and influential, where the harvest is used as a support for human life both as food and a source of income. Many factors can influence the success of agriculture, such as human resources, seed quality and climate that is going on around in the surrounding area. One of the important factors is which climate, the accuracy of determining the climate for agriculture will affect the results obtained. The wrong prediction in determining the future climate will cause crop failure due to incompatibility with the type of plant. In this era, many technologies have been able to predict climate, one of which is technology machine learning that has many types and techniques, which machine learning technology has been widely used in predicting many things. This study aims to predict the climate in an area which is intended to determine crop yields based on the Koppen classification, and also the prediction based on several parameters such as temperature, humidity, duration of sun exposure and rainfall. And the results of this study is have a loss of 0.006 and with the MAPE value as an indicator of the percentage error and as an indicator for determining the accuracy of the prediction results, which is 3.29%, which means that it is included in the very accurate category in predicting climate to estimate agricultural yields.

Keywords –Climate Prediction, Long Short Term Memory (LSTM), Recurrent Neural Network (RNN), Koppen Classification

I. INTRODUCTION

In Foodstuff is one of the important things for human life that can be produced from the agricultural process [1]. Based on it, the results of agriculture are one of the fields that are very close and can support the advancement of economic progress, the better the results obtained, the better the level of the economy [2]. To produce good quality crops, there are many aspects that need to be considered, such as from the climate aspect [3]. The climate itself is the weather conditions in an area that exist at a certain time and is influenced by several factors such as temperature and humidity [4].

Judging from these aspects, climate can be categorized into several types such as rainy season ad dry season, which every season will be affected from every aspect such as temperature, sunlight or humidity [5]. Other then that, changing climates can be influenced by changes in rainfall and also ambient temperature that occurs over a certain period of time [6].

The process of climate forecasting is a difficult thing to do in view of the uncertain climate change. This makes the yields that will be obtained at a certain time period difficult to know. Incorrect climate forecasts in an area can cause considerable losses in agriculture, such as decreasing crop quality, crop failure, decreasing the number of crops and damaged crops [7].

In this era, there have been many technologies capable of predicting the climate in agriculture, one of which is mechine learning technology [8]. Mechine learning technology is one of the mechine lessons that is intended so that a mechine can have the ability to predict and analyze and recognize a pattern [9]. Mechine learning itself has many methods such as K-means, Recurrent Neural Networks, Decision Trees, Artificial Neural Networks, etc [10].

Several previous studies have examined the mechine learning technology used to predict climate. Mechine learning was used to predict maize and soybean yields using the Convolutional Neural Network and Recurrent Neural Network methods. The parameters used are weather, performance, agricultural and soil management [11].

Previous research has also studied mechine learning used to predict climate to be able to see crop yields with parameters temperature, rainfall, cloud cover, humidity and the method used, namely the Decision Tree [12]. Several other studies have also made predictions using the Recurrent Neural Network method which is used to predict bad weather [13], the temperature for the surrounding area [14], wind velocity [15], temperature for daily [16] and rain [17].

There have been many previous studies that have examined the use of technology learning to predict climate in determining agricultural yields. However, there has not been any research that examines the use of mechine learning using the RNN method with Long Short Term Memory and the application of several parameters such as humidity, temperature, rainfall and solar radiation combined with the use of the Koppen classification in classifying climate types. RNN itself is a learning mechine that conducts learning by reviewing previous information. Whereas LSTM is a type of Recurrent Neural Network method that is able to overcome the weaknesses of RNN so that it can improve performance and increase accuracy [18]. On this basis, this method is used, in which climatic conditions are closely related to climatic conditions in the past [19].

The purpose of this research is to predict the climate which is intended to see the yield from agriculture according to Koppen's classification using the RNN LSTM method using temperature, humidity, rainfall and solar radiation as parameters.

II. RESEARCH METHODOLOGY

The object of this research is the climate in the city of Bandung, West Java, the results of rice farming and the Recurrent Neural Network method for the prediction process. In this study, there are several steps that will be carried out in predicting the climate to determine agricultural yields. The steps taken in predicting climate in this study can be seen in Figure 1.

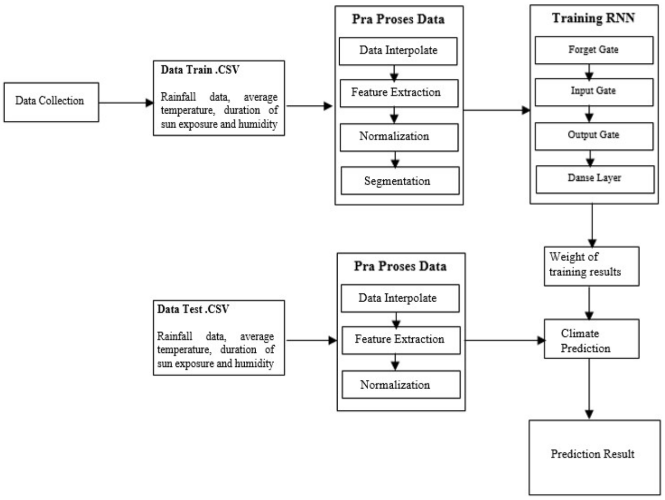


Figure 1. Predicting Climate Step

A. Data Collection

The selection of data in this study is based on the use of Koppen theory which will be used to categorize climate types. The data used in this study are rainfall data, average temperature, duration of sun exposure and humidity for the last 10 years from 2010 to 2020. The data is obtained from the daily climate data of BMKG Bandung City, West Java Province, Bandung Geophysical Station through the official website. BMKG in file format (.csv). This data is daily data, so the amount of data obtained is 3650 x 4 climate parameters = 14600 data

B. Data Training

The training data process will be carried out using the RNN LSTM method which was previously carried out data processing or called data preprocessing, which aims to make the data used for training better and appropriate.

1. Preprocessing Training Data

In this study, the preprocessing was carried out in four stages, namely interpolation, feature extraction, segmentation, and normalization.

- Data Interpolation

Interpolation is the process of finding the value between several known data points. The purpose of interpolation is to correct data that is not measured or not recorded by BMKG by finding the middle value between two values.

- Feature Extraction

Feature extraction is a process to find the largest value in each variable, so that the data will become monthly data with the maximum value of each variable.

- Normalization

Normalization is the process of converting data into normal form. Normalization is needed when the data is very large, very small, or has different units. In this process, the Min-Max normalization is carried out by scaling in the range of zero to one [20].

$$Z = (x - \min(\,))/(\max(\,) - \min(\,)) \qquad (1)$$

X - Data to be normalized,
max - The highest data in the column,
min - The lowest data in the column

- Segmentation

Segmentation is the process of separating and grouping data from raw data into data needed by the system. In this study, data that has become monthly data will be grouped into one year or 12 months with an overlap process, where the one and the next training data has a difference of one month. For example, there are 120 months of climate parameter data, then the distribution of data starts from the 1st month to the 12th month which is used as the first training data, the second month to the 13th month as the second training data, the third month to the 14th month as the third training data, and so on until the 109th training data, namely the 112th month to the 120th month

2. RNN LSTM Training

In the data normalization stage, 109 sets of data were produced to be used as input for the training process using Recurrent Neural Networks. Each data set has 48 total data obtained from 12 months x 4 variables, so that the neurons for input are 36 units. These neurons are connected to neurons contained in the hidden layer, where in the hidden layer there are cells, this process is what distinguishes between ordinary RNN and LSTM. In this cell, there are several steps to be able to produce output.

C. Climate Prediction

At this prediction stage, the MLP architecture will be used in the Recurrent Neural Networks method to predict climate, with the input data used in the input layer, namely rainfall, average temperature, duration of sun exposure and humidity. The MLP architecture designed in the Recurrent Neural Networks method can be seen in Figure 2.

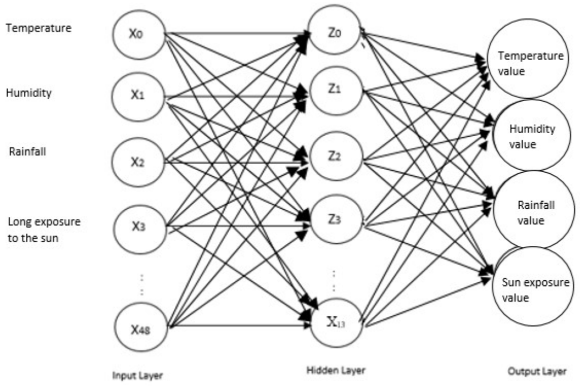
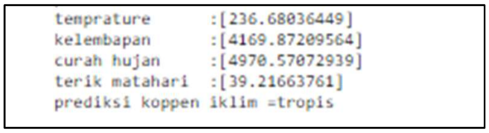


Figure 2. MLP Architecture for Climate Prediction

As can be seen in Figure 2 regarding the MLP architecture used in this study to predict climate, where there are 48 neurons in the input layer obtained from 12 months x 4 climate parameters. There are 13 neurons in the hidden layer which are calculated using equation 2.9 and in the output layer there is one neuron that represents the output of the predictions. The testing process uses the weight of the training results that have been stored in the form of a file, and the testing process will produce an output value in the form of a climate prediction value for each variable. Then the climate value will be entered into the range of values that have been determined using the Koppen Classification

D. *Prediction Result*

This study will produce a prediction of the climate based on four parameters such as rainfall data, average temperature, duration of sun exposure and humidity for the next one month and then the prediction results will be categorized into climate types based on Koppen's theory. The prediction result can be seen in Figure 3.



```
temprature       :[236.68036449]
kelembapan       :[4169.87209564]
curah hujan      :[4970.57072939]
terik matahari   :[39.21663761]
prediksi koppen iklim =tropis
```

Figure 3. MLP Architecture for Climate Prediction

III. Results And Discussion

The test consisted of the method effect test, namely between the optimization model effect test, and the test for the effect of the number of epochs. At this stage an analysis of each tested parameter is carried out.

Before test it,, the training process has been carried out using data that has passed the data preprocess.

Table 2. Data training for first batch

Training Data								
Data - 1				**...**	**Data – 12**			
SU	**KL**	**CH**	**LP**	**...**	**SU**	**KL**	**CH**	**LP**
24,6	91	86	8	...	24	92	78	6,7
24,9	93	82,4	6,6	...	24,4	85	17,5	7,4
24,6	94	94	7,5	...	25,2	88	20,5	8
26	85	27	8	...	25,2	90	73,5	7,4
25,4	94	92	7,7	...	25,3	86	56	7,9
24,9	94	27,4	7,9	...	24,3	87	44,4	8
24,9	94	61	7,7	...	24,3	84	36,9	**8**
...	...	...	...	...	...	...	...	...

SU - Temperature
KL - Humidity
CH - Rainfall
LP - Duration of sun exposure

The training process will be carried out using the data in the Table 2 by using the RNN LSTM architecture. In addition, training is also carried out using two optimizers, namely Adam and Sigmod with several epoch including 50, 100, 200 and also used 0,001 for learning rate.

A. *Optimization Model Influence Test*

The optimization model effect test is conducted to determine which optimization model is more suitable for predicting climate. This optimization model is used to update the weights during training using two optimization models, namely Adam optimizer and SGD optimizer. The accuracy results from testing the two optimization models can be seen in Table 3.

Table 3. Test the Influence of the Optimization Model

No	Optimizer	Loss	MAPE(%)
1	Adam	0,006	32,9
2	SGD	0,011	13,1

After testing, it was found that the Adam optimization method has a lower loss than SGD. Graph of SGD and Adam test results can be seen in Figure 3 and Figure 4.

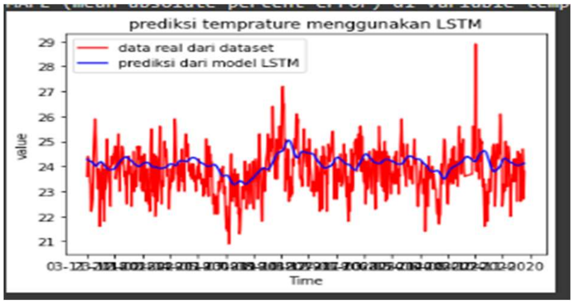


Figure 4. Results of SGD Loss and MAPE

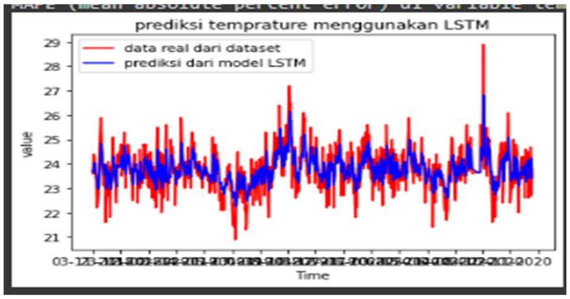


Figure 5. Adam's Loss and MAPE Results

It is found that Adam's optimization has a lower loss than SGD optimization, which is 0.006 and with MAPE measurement as an indicator of the percentage error of the prediction results, which is 3.29%, which means that it is included in the very accurate category. Meanwhile, SGD itself has a loss of 0.011 and MAPE of 13.1%, which means it is in the good category

B. Test of the Influence of the Number of Epochs

This study used 200 epochs with low loss results. The results of the epoch test can be seen in Table 4.

Table. 4. Test for the Effect of the Number of Epochs

No	Epoch	Loss	MAPE(%)
1	50	0,009	7,83
2	100	0,007	3,29
3	200	0,006	2,36

It was found that the use of epoch 50 has the highest loss value with a value of 0.009 and a MAPE of 7.83%. Meanwhile, the epoch which had the lowest loss value was 200 and MAPE 2.36%.

IV. Conclusion

This research produces climate prediction using RNN LSTM. This prediction can provide results in the form of climate forecasts for the next day. The results of this prediction are in the form of climate types belonging to tropical climates, dry climates, temperate climates, continental climates and arctic climates. This prediction can be used by the user in predicting the future climate as a benchmark to determine which farm will produce good yields for that climate. Climate prediction in this study produces low loss using Adam's optimization and learning rate of 0.001 and the number of epochs is 200, which results in a loss of 0.006 and with the MAPE value as an indicator of the amount of error percentage and as an indicator of determining the accuracy of the prediction results, which is 3.29%. means it falls into the very accurate category.

References

[1] Abbas, Z., Al-Shishtawy, A., Girdzijauskas, S., & Vlassov, V. (2018). Short-Term Traffic Prediction Using Long Short-Term Memory Neural Networks. *2018 IEEE International Congress on Big Data (BigData Congress)*, 57-65.

[2] Adedeji, O., Reuben, O., & Olatoye, O. (2014). Global Climate Change. *Journal of Geoscience and Environment Protection*(2), 114-122.

[3] Beck, H., Zimmermann, N. E., McVicar, T., & Vergopolan, N. (2018). Present and Future Köppen-Geiger climate Classification Maps at 1-Km Resolution.

[4] Chenn, D., & Chen, H. W. (2013). Using the Koppen Classification to Quantify Climate Variation and Change: An Example for1901–2010. *Environmental Development*, 63-79.

[5] Rippke, U., Ramirez-Villegas, J., Jarvis, A., & Vermeulen, S. J. (2016). Timescales of transformational climate change adaptation in Sub-Saharan African agriculture . *Nature Climate Change*.

[6] Feng, Q. Y., Vasile, R., Segond, M., Gozolchiani, A., Wang, Y., Abel, M., . . . Dijkstra1, H. A. (2016). ClimateLearn: A Machine-Learning Approach for Climate Prediction Using Network Measures. *Geosci Model Dev Discuss*.

[7] Salman, A. G., Heryadi, Y., Abdurahman, E., & Suparta, W. (2018). Weather Forecasting Using Merged Long Short-term Memory Model . *Bulletin of Electrical Engineering and Informatics, 7*(3), 377-385

[8] Fente, D. N., & Singh, D. K. (2018). Weather Forecasting Using Artificial Neural Network . *2nd International Conference on Inventive Communication and Computational Technologies (ICICCT)*.

[9] Huang, Y. (2019). A Prediction Scheme for Daily Maximum and Minimum Temperature Forecasts Using Recurrent Neural Network and Rough set. *IOP Conference Series: Earth and Environmental Science ICAESEE*.

[10] Jahan, I., Sajal, S. Z., & Nygard, K. E. (2019). Prediction Model Using Recurrent Neural Networks . *IEEE International Conference on Electro Information Technology (EIT)*.

[11] Jin, J., Li, M., & Jin, L. (2015). Data Normalization to Accelerate Training for Linear Neural Net to Predict Tropical Cyclone Tracks. *Mathematical Problems in Engineering*, 8.

[12] Kaunang, F. J., Rotikan, R., & Tulung, G. S. (2018). Pemodelan Sistem Prediksi Tanaman Pangan Menggunakan Algoritma Decision Tree . *Cogito Smart Journal, vi*(1).

[13] Khaki, S., Wang, L., & Archontoulis, S. V. (2019). A CNN-RNN Framework for Crop Yield Prediction. *Frontiers in Plant Science, x*.

[14] Kumar, N., Kaur, G., & Aditi. (2017). Wind Speed Prediction using Neural Network . *International Journal of Advanced Production and Industrial Engineering IJAPIE-SI-IDCM , 608*, 36-41.

[15] Liu, Y., Wang, Y., Yang, X., & Zhang, L. (2017). Short-Term Travel Time Prediction by Deep Learning: A Comparison of Different LSTM-DNN Models. *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*.

[16] Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., & Kedam, G. (2019). A Machine Learning Model for Air Quality Prediction for Smart Cities. *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*.

[17] Mubyarto. (1989). *Ekonomi pertanian.* Jakarta: LP3ES.

[18] Naveen Kumar Arora. (2019). Impact of Climate Change on Agriculture Production and Its Sustainable Solutions. *Environmental Sustainabilit*(2), 95-96.

[19] Nilsson, & J, N. (2015). *Introduction to Mechine Learning.* Standford University.

[20] Nilsson, N. J. (2015). *Introduction to Machine Learning.*

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Features Selection based on Enhanced KNN to Predict Raw Material Needs on PT. SANM

Siti Aisyah Naili Mutia[1*)], Tjong Wan Sen[2]
[1]Program Studi Teknik Informatika, Fakultas Komputer, President University
[2]Program Studi Teknik Informatika, Fakultas Komputer, President University
Email: [1]sitiaisyah.mutia@student.president.ac.id, [2]wansen@president.ac.id

Abstract – Raw material inventory must be able to meet production needs. So it is necessary to plan / predict raw material needs in the following month to determine the raw material inventory. Currently PT. SANM uses a manual counting method, the expenditure of raw materials for six months, then deducts the current raw material inventory. As a result, there are raw materials that are over order or lacking, which causes production to be constrained. The manual calculation method is not effective enough to meet the raw material inventory. In this research, the researcher proposes an algorithm which is contained in Data Mining, that is Enhanced KNN using GWO to predict raw material needs. Because GWO and Enhanced KNN algorithms give the results are easy to understand, have good accuracy compared to other machine learning methods, can cover the trapped problem from KNN traditional and capable of improving the accuracy using feature selection method. The method used in this study is to compare Enhanced KNN with and without GWO that gives a significant increase in the accuracy value by 16.5%, from 44.6% to 61.1%.

Keywords – Data mining, Classification, Enhanced K-Nearest Neighbor, Feature Selection

I. INTRODUCTION

PT. SANM is a company that operates in the field of eyeglass lenses, which includes medical devices. There are many types of raw materials, ranging from different segments, different indexes, white or color and diameter. According to the type of segment, the lens can be divided into three, namely a circular segment or called a kryptok lens, a semicircular segment or called a flattop and does not have a segment called single vision. According to the index there are index lenses 1.50, 1.53, 1.56, 1.57, 1.60, 1.67, 1.74. According to the color there are white, photochromic gray, photochromic brown, nupolar green, nupolar gray and also anti-radiation. Meanwhile, according to the diameter there are lenses with a diameter of 60, 65, 70, 72, 74, 76.

PT. SANM only carries out production according to orders or what is usually called a business process driven by order. When a customer places an order with various types of lenses, spherical size, cylinder or add, then the production process will be carried out. So that there is no production planning. Currently, to order raw materials, the manual calculation method is still used. The raw material for this lens does not have an expiration date, but if it is stored in an inappropriate manner and time, the raw material can no longer be used because it is out of standard or damaged.

To avoid the problem of shortage or excess of raw materials, it is necessary to classify the prediction of raw material needs. One of the classification algorithms that can be used is KNN[1].

Estimates for predicting disturbances from past data/events in the industrial sector using KNN. This study also explains predictions according to supervised data about changes in the characteristics of disorders that have occurred[2]. The classification results using KNN are easy to understand and have good accuracy compared to other machine learning methods[3] or deep learning with newer approach using speech[4] or image[5] as input. The advantage of the KNN method is that it is effectively applied to large amounts of data and is resilient to noise training data, which is data that has the farthest range of values compared to other data but can disrupt the existing data structure. In addition to advantages, KNN also has weaknesses, namely less than optimal in determining the value of k which is the number of nearest neighbors and must determine the attribute to be selected or feature selection in order to get the best results, the presence of outliers in the training data and sensitive to the value of k (there are no clear rules). to determine the optimal value of k). Despite its simplicity and high efficiency, traditional KNNs can easily get stuck[6].

To overcome these shortcomings, an improvement solution is needed to optimize the KNN classification. To overcome these shortcomings, an improvement solution is needed to optimize the KNN classification. One solution that can be applied is to perform optimization by applying EKNN[6] and adding a feature selection process using the GWO algorithm[7][8][9]. Based on the above studies, the researcher chose the EKNN method to classify and predict the raw material needs of the coming month and used GWO as a feature selection method to improve EKNN performance.

II. RESEARCH METHODOLOGY

The research stages are outlined in Figure 1. which describes the research process that will be taken.

In this research, there are several stages carried out, including: Preparation, includes background creation, problem identification, problem formulation, objectives, problem boundaries, research benefits and novelty. Study literature by conducting a literature review on predictions using EKNN. Data collection, namely data collection by collecting sales data for 2018-2020. Details of the dataset used are shown in Table 1.

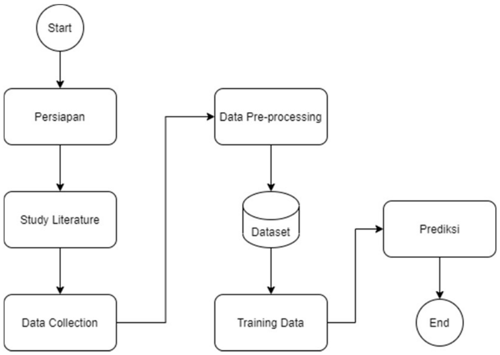


Figure 1 Research Stages

Table 1 Dataset Detail

Dataset		
Data	***Attribute***	***Target***
60	34	3

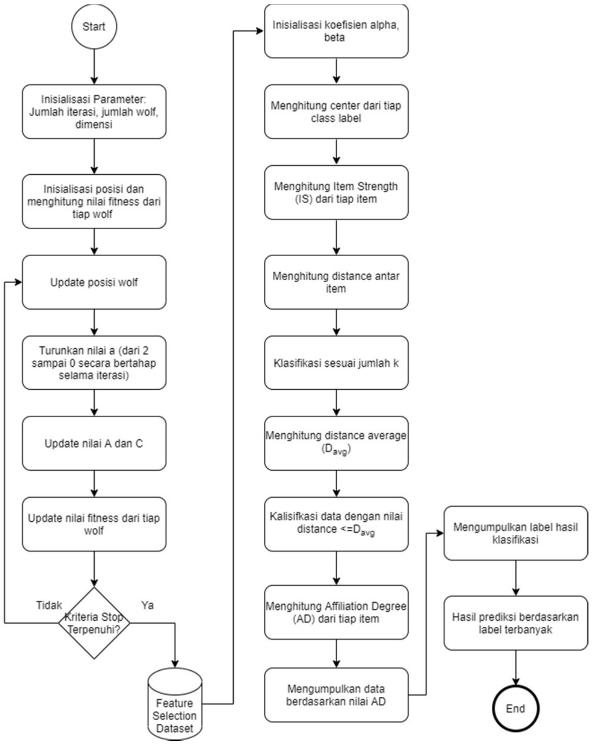

Start
Inisialisasi Parameter: Jumlah iterasi, jumlah wolf, dimensi
Inisialisasi posisi dan menghitung nilai fitness dari tiap wolf
Update posisi wolf
Turunkan nilai a (dari 2 sampai 0 secara bertahap selama iterasi)
Update nilai A dan C
Update nilai fitness dari tiap wolf
Tidak
Kriteria Stop Terpenuhi?
Ya
Feature Selection Dataset
Inisialisasi koefisien alpha, beta
Menghitung center dari tiap class label
Menghitung Item Strength (IS) dari tiap item
Menghitung distance antar item
Klasifikasi sesuai jumlah k
Menghitung distance average (D_{avg})
Kalisifkasi data dengan nilai distance <=D_{avg}
Menghitung Affiliation Degree (AD) dari tiap item
Mengumpulkan data berdasarkan nilai AD
Mengumpulkan label hasil klasifikasi
Hasil prediksi berdasarkan label terbanyak
End

Figure 2 Training Model

Data pre-processing, namely removing data noise such as duplicate data, empty values, transforming data from time series into per material code and data normalization. Data training will be carried out using the EKNN and GWO algorithms. The researcher proposes the EKNN algorithm because it is easy to understand and provides better accuracy results than other machine learning methods[3] and is able to overcome the problem of being trapped in traditional KNN[6]. Then the use of GWO as a feature selection method to increase the accuracy value on EKNN. The experiment in the first step will be trained using the EKNN algorithm and then using EKNN-GWO to get a comparison of the results of accuracy, precision and recall values, as shown in Figure 2.

A. *Enhanced KNN*

The Enhanced KNN algorithm is a development algorithm from K-Nearest Neighbor. Traditional KNN is easily trapped in several situations, such as there are two classes that have the same number of neighbor classifications/series. Examples of traps that may occur in traditional KNN can be seen in Figure 3.

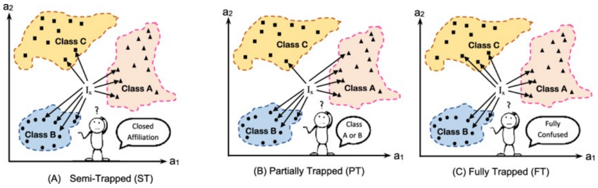


Figure 3 Different-types-of-KNN-trapping [6]

To overcome this weakness, another parameter is needed to determine the closest neighbor. Like the traditional KNN, the EKNN deals with its closest neighbors with the item being tested in the feature room.

```
Input: Training dataset, center, α, β
Output: Item Strength (IS)
Steps:
For each class label (cx € CL) do:
        Project the training examples in the n dimensional feature
      space
Next
For each class (cx € CL) do:
For each item (Ij € cx) do:
        Calculate the strength of item Ij as:
            $IS(I_j) = \frac{\left[\alpha * \sum_{j \neq k} \frac{1}{Dis(I_j, I_k)} + \frac{\beta}{Dis(I_j, C)}\right]}{2}$
    Next
Next
```

Figure 4 EKNN Phase Training Algorithm

However, these neighbors will have different strengths according to the level of closeness to the item being tested and the neighbor's membership level with their class, which is called Item Strength (IS). The EKNN is implemented in two stages, namely; the training and testing phase[6].

In the training phase, all strength is counted. Where item strength is a measure for the relationship level of the item with its hosting class. Assuming *n* features, *m* as the target class, to calculate the strength of each item. EKNN training algorithm can be seen in Figure 4 and EKNN training phase flow can be seen in Figure 5.

First, all items are projected into the dimension space feature. Then, the center of each class containing examples in n dimensionless feature spaces can be solved using equation (1).

$$C = \left\{ \frac{\sum_{q=1}^{t} V_q^1}{t}, \frac{\sum_{q=1}^{t} V_q^2}{t}, \ldots\ldots\ldots, \frac{\sum_{q=1}^{t} V_q^n}{t} \right\} \qquad (1)$$

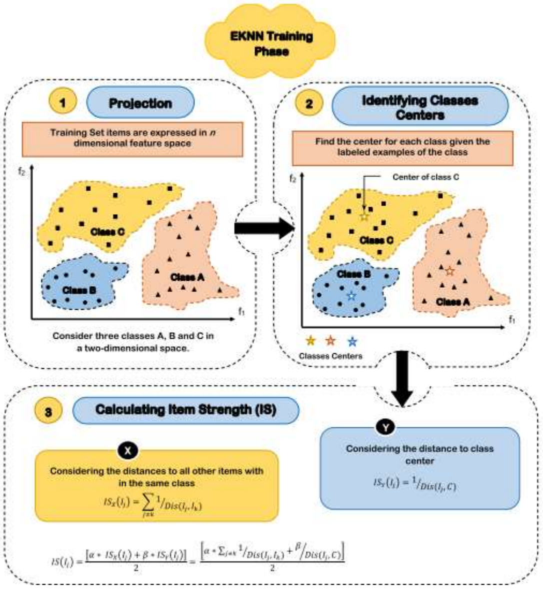

EKNN Training Phase
1
Projection
Training Set items are expressed in n dimensional feature space
f2
Class C
Class B
Class A
f1
Consider three classes A, B and C in a two-dimensional space.
2
Identifying Classes Centers
Find the center for each class given the labeled examples of the class
Center of class C
Classes Centers
3
Calculating Item Strength (IS)
X
Considering the distances to all other items with in the same class
$IS_X(I_j) = \sum_{j\neq k} {}^1/_{Dis(I_j,I_k)}$
Y
Considering the distance to class center
$IS_Y(I_j) = {}^1/_{Dis(I_j,C)}$
$IS(I_j) = \frac{[\alpha * IS_X(I_j) + \beta * IS_Y(I_j)]}{2} = \frac{\left[\alpha * \sum_{j\neq k} {}^1/_{Dis(I_j,I_k)} + {}^\beta/_{Dis(I_j,C)}\right]}{2}$

Figure 5 EKNN Training Phase [6]

Where C is the class center in the n-dimensional feature space, t is the number of instances in the class, and V_q^i is the i-th dimension value of the q-th example. Then, the strength of item I_j can be calculated using equation (2).

$$IS(I_j) = \frac{[\alpha * IS_X(I_j) + \beta * IS_Y(I_j)]}{2} \qquad (2)$$

Where $IS(I_j)$ is the power of the item I_j, which is the weighted average of the two values. The first measures the strength of an item based on its proximity to another item (example) in its class and is denoted as; $IS_X(I_j)$. On the other hand, the value further considers the proximity of the item to the center of the class as an indication of its degree of affiliation with the class and according to its strength, which is denoted as; $IS_Y(I_j)$. As in the previous equation, α and β are weighting factors that express the relative impact of $IS_X(I_j)$ and $IS_Y(I_j)$, where $0<\alpha\leq 1$ and $0<\beta\leq 1$. In general, $IS_X(I_j)$ and $IS_Y(I_j)$ can be calculated using equations (3) and (4), respectively.

$$IS_X(I_j) = \sum_{j\neq k} \frac{1}{Dis(I_j,I_k)} \qquad (3)$$

$$IS_Y(I_j) = \frac{1}{Dis(I_j,C)} \qquad (4)$$

Where $IS_X(I_j)$ and $IS_Y(I_j)$ are the strength of the item considering the closeness of each class item and class center.

$Dis\ (I_j,\ I_k)$ is the Euclidean distance between item I_j and I_k, and $Dis(I_j,\ C)$ is the Euclidean distance between item I_j and the class center. Calculating the distance between two points p_x and p_y in a feature space of dimension n can be calculated using equation (5).

$$Dis(p_x, p_y) = \sqrt{\sum_{i=1}^{n} \left(p_x^i - p_y^i\right)^2} \qquad (5)$$

Where $p_x i$ and p_y^i are the i-th dimensional values of the p_x and p_y points in the n-dimensional feature space, respectively. Then, in place of (2), the power of item I_j can be calculated using equation (6).

$$IS(I_j) = \frac{\left[\alpha * \sum_{j\neq k} \frac{1}{Dis(I_j,I_k)} + \frac{\beta}{Dis(I_j,C)}\right]}{2} \qquad (6)$$

An important issue is how to estimate the optimal values of α and β. As tunable parameters, the optimal values of α and β can be calculated empirically by assigning different values in a predetermined scenario, then calculating the resultant accuracy of the EKNN classifier by considering a series of test items. The optimal α and β values are those that provide maximum classification accuracy. The suggested scenario is to start with initial values of α and β, for example $\alpha = 0$ and $\beta = 0$, the values of α and β will be increased using a constant positive step ξ keeping the values of α and β greater than 0 and less than or equal to 1.

In the testing phase, the proposed EKNN, consider ring a target class set m to CL = {c 1, c 2, ... cm}. EKNN testing algorithm can be seen in Figure 6 and EKNN testing phase flow can be seen in Figure 7.

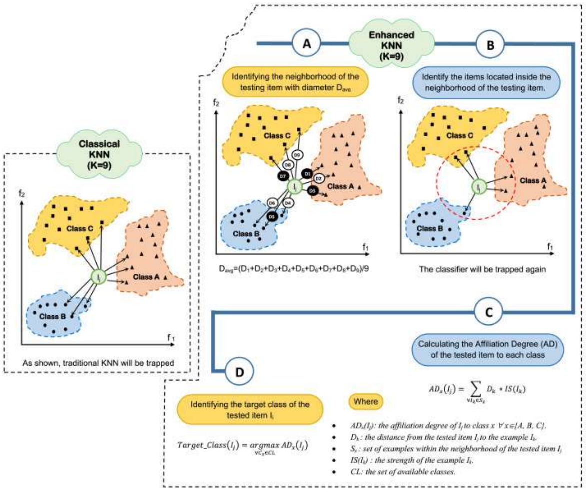
Initially, the item under test is projected in a feature of n dimensional space. Then, the closest K-neighbor was identified. The distance from the tested item Ij to each k nearest neighbor is calculated, then the average *distance(D_{avg})* is then calculated. Circular spheres are identified whose radius is equal to D_{avg}. Only samples in the identified environment will be considered for classifying new items. Affiliation Degree (AD) items tested for each class are calculated using the equation (7).

$$AD_x(I_j) = \sum_{\forall I_k \in S_x} D_k * IS\ (I_k) \qquad (7)$$

Input: Training dataset, IS
Output: Classifying the tested item Ij to one of the available target class
Steps:

1. The project the tested item into the n dimensional feature space.
2. Pick the nearest K items near the test point Ij.
3. Computer the average distance Davg from Ij and each of the K nearest examples
4. Pick items located in the circle with diameter Davg
5. Calculate the Affiliation Degree (AD) of the tested items to each classes;
$$AD_x(I_j) = \sum_{\forall I_k \in S_x} D_k * IS\ (I_k)$$
6. Identify the target class of the tested item;
$$Target\_Class(I_j) = \underset{\forall C_x \in CL}{argmax} AD_x\ (I_j)$$

Figure 6 EKNN Algorithm Testing Phase


Classical KNN (K=9)
f2
Class C
Class A
Class B
f1
As shown, traditional KNN will be trapped
Enhanced KNN (K=9)
A
B
Identifying the neighborhood of the testing item with diameter D_avg
Identify the items located inside the neighborhood of the testing item.
D1
D2
D3
D4
D5
D6
D7
D8
D9
Ij
D_avg=(D1+D2+D3+D4+D5+D6+D7+D8+D9)/9
The classifier will be trapped again
C
Calculating the Affiliation Degree (AD) of the tested item to each class
D
Identifying the target class of the tested item Ij
Where

$$AD_x(I_j) = \sum_{\forall I_k \in S_x} D_k * IS(I_k)$$

$$Target\_Class(I_j) = \underset{\forall C_x \in CL}{argmax}\, AD_x(I_j)$$

- $AD_x(I_j)$: the affiliation degree of I_j to class x $\forall x \in \{A, B, C\}$.
- D_k: the distance from the tested item I_j to the example I_k.
- S_x: set of examples within the neighborhood of the tested item I_j
- $IS(I_k)$: the strength of the example I_k.
- CL: the set of available classes.

Figure 7 EKNN Testing Phase [6]

Where AD_xI_j is the level of affiliation of item I_j to tested class x∀ x ∈{A B C}. D_k is the distance from the tested problem I_j to the example I_k. S_x is the set of examples in the environment of the items tested I_j. And $IS(I_k)$ is the strength of the I_k example. Finally, the item under test is targeted to the class that has the maximum level of affiliation as illustrated in equation (8).

$$Target\_Class(Ij) =_{\forall CxCL}^{argmaxADx(Ij)} \quad (8)$$

Where CL is the set of the target class under consideration.

B. Feature Selection

Feature selection is an optimization problem that plays an important role in dealing with classification problems. It is the process of selecting an optimal subset of features from a data set so that the classifier can obtain better accuracy and/or reduced computational load. However, removing irrelevant features is challenging and time consuming due to the large search space and the relationship between features[10]. The traditional FS method has the disadvantage of nesting effects and computational costs. To solve this problem, population-based optimization algorithms, such as gray wolf optimization (GWO)[8], particle swarm optimization (PSO)[11], genetic algorithm (GA), genetic programming (GP), ant colony optimization (ACO), brain storm optimization (BSO) are used. and harmony search (HS)[12].

C. Grey Wolf Optimization (GWO)

Gray Wolf Optimization (GWO) is a population-based metaheuristic algorithm that simulates leadership hierarchies and the mechanism of hunting grey wolves in nature discovered by Mirjalili in 2014[8]. The gray wolf (Gray Wolf) is a carnivorous animal that is at the highest level in the food chain, otherwise known as the Apex Predator. Gray wolves live in groups, where each group consists of 5-12 individuals[7]. The interesting thing about gray wolves is their social level. The gray wolf is divided into four social levels, as in Figure 8.

The first level is called Alpha (α), which is the leader of a group where they are female and male wolves. The first level wolf has the most right to make all decisions, regarding hunting activities, hunting time, resting place and so on.

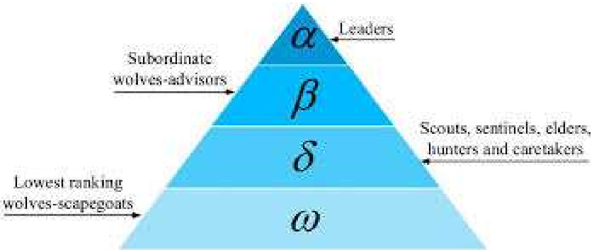

α
Leaders
Subordinate wolves-advisors
β
Scouts, sentinels, elders, hunters and caretakers
δ
Lowest ranking wolves-scapegoats
ω

Figure 8 Gray Wolf Level [8]

The wolf's orders of the first level will be obeyed by all wolves that are in the lower levels. The second level wolf is also called Beta (β), which is in charge of assisting the first level wolf in making decisions, as an advisor and as a substitute for the first level wolf when it dies. The second level wolf can be male or female. The third level wolf or Delta (δ) is usually called subordinate. The third level wolf has several categories; scouts are in charge of guarding the territory and giving warnings when in danger, sentinels as group protectors, elders who are experienced and have the potential to replace first and second level wolves, hunters who help first and second level wolves in hunting and provide food, and caretakers as carers for a wolf that is sick or injured.ω

Apart from the social level, another interesting thing about the gray wolf is group hunting. At the first level GWO is considered the best solution.

The steps of hunting and mathematical models on GWO are as follows:

Step 1: Encircling prey

$$\vec{D} = \left|\vec{C} \times \overrightarrow{X_p}(t) \ - \vec{X}(t)\right| \quad (9)$$

$$\vec{X}(t+1) = \overrightarrow{X_p}(t) \ - \vec{A} \times \vec{D} \quad (10)$$

Where t denotes the current iteration, $\vec{A}$ and $\vec{C}$ is the vector coefficient, $\overrightarrow{X_p}$ is the vector of the prey position, and $\vec{X}$ denotes the vector of the position of the gray wolf. Where $\vec{A}$ and $\vec{C}$ can be calculated by the equation below.

$$\vec{A} = 2\vec{a} \times \vec{r}_1 - \vec{a} \quad (11)$$

$$\vec{C} = 2 \times \vec{r}_2 \quad (12)$$

Where the component is derived linearly from 2 to 0 during iteration and r1, r2 are random vectors in [0,1].$\vec{a}$

Step 2: Hunting: hunting activities are usually guided by Alpha level wolves. Beta and Delta might participate in the hunt every now and then. In the mathematical model of the gray wolf hunting behavior, it is assumed that Alpha, Beta, and Delta have better knowledge of the potential location of the prey. The first three best solutions are saved and the other wolves must update

e-ISSN: 2614-8404
p-ISSN: 2776-3234

their position according to the position of the best seeker wolf as shown in the equation below.

$$\vec{D}_{\propto} = \left|\vec{C}_1 \times \vec{X}_{\propto} - \vec{X}\right|$$
$$\vec{D}_{\beta} = \left|\vec{C}_2 \times \vec{X}_{\beta} - \vec{X}\right|$$
$$\vec{D}_{\delta} = \left|\vec{C}_3 \times \vec{X}_{\delta} - \vec{X}\right| \quad (13)$$

$$\vec{X}_1 = \vec{X}_{\propto} - \vec{A}_1 \times \left(\vec{D}_{\propto}\right),$$
$$\vec{X}_2 = \vec{X}_{\beta} - \vec{A}_2 \times \left(\vec{D}_{\beta}\right)$$
$$\vec{X}_3 = \vec{X}_{\delta} - \vec{A}_3 \times \left(\vec{D}_{\delta}\right) \quad (14)$$

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (15)$$

Step 3: Attacking the prey (exploitation): The gray wolf completes the hunt by attacking its prey when it stops moving. The vector is a random value in the interval [-2a, 2a], where a is lowered from 2 to 0 during iteration. When the wolf attacks the prey, that represents a process of exploitation.$\vec{A}\left|\vec{A}\right| < 1$

Step 4: Looking for prey (exploration): The roaming process in GWO is applied according to different positions and positions to find prey and gather to attack the prey. The exploration process is modeled mathematically by utilizing A with a random value greater than 1 or less than -1 to oblige the seeker group to deviate from its prey. When | A |> 1, the wolf is forced to deviate from its prey to become a fitter prey.

D. Validation

One alternative approach to “train and test” that is often adopted in some cases (and some regardless of size) is called K-Fold Cross Validation.[13] by testing the amount of error in the test data[14]. We use the k-1 sample for training and the remaining 1 sample for testing. For example, there are 10 data subsets, we use 9 subsets for training and the remaining 1 subset for testing. There are 10 training times where in each training there are 9 data subsets for training and 1 subset is used for testing. After that, the average error and standard deviation of error are calculated[14]. Each k part is in turn used as the test set and the other k -1 parts are used as the training set[15].

E. Measurement Stages

Confusion matrix is a dataset having only two classes, one class as positive and the other class as negative[15]. This method uses a configuration matrix model table.

Table 2 Confusion Matrix Model[16]

		Prediction		**Total**
		-1 (Negative)	**+1 (Positive)**	
Examples	**-1 (Negative)**	p	q	p + q
	+1 (Positive)	u	v	u + v
	Total	p + u	q + v	m

Where p is the number of correct predictions that the instance will be negative. q is the number of incorrect predictions that the instance will be positive. u is the number of incorrect predictions that the instance will be negative. v is the number of correct predictions that the instance will be positive.

Here is the confusion matrix model equation:

a. The accuracy value (acc) is the proportion of the number of correct predictions. Can be calculated using the equation:

$$acc = \frac{p+v}{(p+q+u+v)} = \frac{p+v}{m} \quad [9]$$

b. The recall or true positive rate (tp) is the proportion of correctly classified positive cases calculated using the equation:

$$tp = \frac{u}{u+v} \quad [10]$$

c. Precision (p) is the proportion of correct positive case predictions, which are calculated using the equation:

$$p = \frac{v}{q+v} \quad [11]$$

III. Results And Discussion

Data collection was obtained from PT SANM. The dataset used in this study is sales data for the years 2018-2020. The data contains 60 rows and 34 attributes among other BL01, BL01BB, BL01P, BL01PUV, BL02, BL02BB, BL02PUV, BL05, BL05BB, BL05BB1.6, BL05BL, BL05D, BL05DW, BL05L, BL05LB, BL05M, BL05M1.6, BL05M1.67, BL05NU, BL05NUG, BL05O, BL05OB, BL05PG, BL05PGUV, BL05PGUV1.6, BL05PGUV1.67, BL05SSS, BL05YVII, BL05YVII1.67, BL06PB, IMPACT1.57, SV1.67, SV1.74 and TRIVE1.53. The data pre-processing stage used in this study is to set aside unused attributes and data transformation. Data transformation is used to change the number class data type to nominal. This research classification method uses Enhanced KNN based[6].

One alternative approach to “train and test” that is often adopted in some cases (and some regardless of size) is called K-Fold Cross Validation.[17]. The results of the accuracy of each fold are shown in Table 3.

Table 3 Test Result Accuracy Value

Fold	EKNN	GWO-EKNN
1	33%	83%
2	50%	33%
3	83%	33%
4	50%	83%
5	50%	0%
6	50%	33%
7	33%	83%

8	17%	83%
9	40%	100%
10	40%	80%
Average	44.6%	61.1%

From 10 tests, the average accuracy of the EKNN method is 44.6%, with the highest accuracy value of 83% and the lowest accuracy value of 17%. Then the test was carried out using the GWO-EKNN method with an average accuracy value of 61.1%, with the highest accuracy value of 100% and the lowest accuracy value of 0%.

Based on the accuracy value data between the EKNN and GWO-EKNN methods, the results show that the GWO-EKNN accuracy value is better than the EKNN with an average accuracy value of 61.1% with a difference of 16.5%. The following is a graphical comparison result to see a clearer difference in accuracy values.

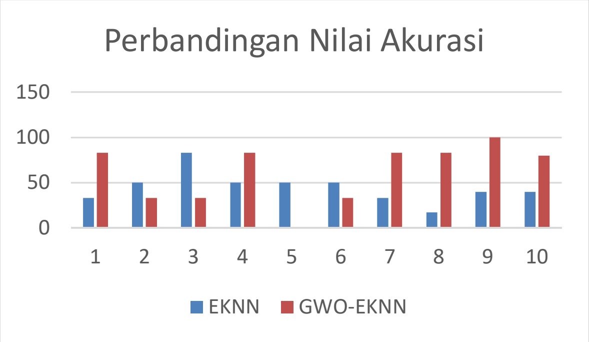


Figure 9 Comparison of the accuracy value of EKNN with GWO-EKNN

In Figure 9, the highest average accuracy value is found in the GWO-EKNN algorithm. In the experiments that have been carried out, it can be concluded that the experimental results are in the table below:

Table 4 Comparison of accuracy, precision and recall values

Method	Average Accuracy	Average Precision	Average recall
EKNN	44.6%	23.3%	44.4%
GWO-EKNN	**61.1%**	**52.5%**	**62.8%**

Based on Table 4, it shows that the best experimental results are in the GWO-EKNN algorithm with an accuracy value of 61.1%, a precision value of 52.5% and a recall value of 62.8%. The second method is the EKNN algorithm with an accuracy value of 44.6%, a precision value of 23.3% and a recall value of 44.4%.

IV. Conclusion

Based on the research that has been done to improve the performance of EKNN in predicting raw material needs by selecting features using GWO, it can be concluded that GWO is able to increase EKNN performance by 16.5%, from 44.6% to 61.1%. Based on the GWO feature selection method, it was found that the features that can increase the accuracy of the prediction of raw material requirements include BL01, BL01BB, BL01P, BL01PUV, BL02, BL02PUV, BL05BB, BL05BB1.6, BL05BL, BL05DW and Target. So that the GWO-EKNN method can predict the raw material needs more precisely on the sales dataset of PT. SANM in 2018-2020.

Based on the above conclusions, for future research it is recommended to use other metaheuristic feature selection methods, such as: PSO, GA, ACO, BCO, and others as well as using optimization of metaheuristic methods.

References

[1] R. Arian, A. Hariri, A. Mehridehnavi, A. Fassihi, and F. Ghasemi, “Protein kinase inhibitors’ classification using K-Nearest neighbor algorithm,” *Comput. Biol. Chem.*, vol. 86, no. April, p. 107269, 2020, doi: 10.1016/j.compbiolchem.2020.107269.

[2] F. Borghesan, M. Chioua, and N. F. Thornhill, “Forecasting of process disturbances using k-nearest neighbours, with an application in process control,” *Comput. Chem. Eng.*, vol. 128, no. 675215, pp. 188–200, 2019, doi: 10.1016/j.compchemeng.2019.05.009.

[3] D. A. Adeniyi, Z. Wei, and Y. Yongquan, “Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method,” *Appl. Comput. Informatics*, vol. 12, no. 1, pp. 90–108, 2016, doi: 10.1016/j.aci.2014.10.001.

[4] T. W. Sen, “Voice Activity Detector for Device with Small Processor and Memory,” *ICSECC 2019 - Int. Conf. Sustain. Eng. Creat. Comput. New Idea, New Innov. Proc.*, pp. 212–217, 2019, doi: 10.1109/ICSECC.2019.8907081.

[5] T. W. Sen, S. Suakanto, A. M. Siregar, and A. L. P. Localization, “License Plate Localization for Low Computation Resources Systems Using Raw Image Input and Artificial Neural Network,” vol. 15, no. 1, pp. 39–44, 1858.

[6] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsoud, “A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier,” *Knowledge-Based Syst.*, vol. 205, p. 106270, 2020, doi: 10.1016/j.knosys.2020.106270.

[7] M. A. Oktaviani, R. S. Wibowo, and N. K. Aryani, “Aliran Daya Optimal dengan Efek Katup Menggunakan Grey Wolf Optimization,” *J. Tek. ITS*, vol. 7, no. 2, 2019, doi: 10.12962/j23373539.v7i2.30906.

[8] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey Wolf Optimizer,” *Adv. Eng. Softw.*, vol. 69, pp. 46–61, 2014, doi: 10.1016/j.advengsoft.2013.12.007.

[9] M. H. Nadimi-Shahraki, S. Taghian, and S. Mirjalili, “An improved grey wolf optimizer for solving engineering problems,” *Expert Syst. Appl.*, vol. 166, p. 113917, 2021, doi: 10.1016/j.eswa.2020.113917.

[10] J. Gholami, F. Pourpanah, and X. Wang, "Feature selection based on improved binary global harmony search for data classification," *Appl. Soft Comput. J.*, vol. 93, p. 106402, 2020, doi: 10.1016/j.asoc.2020.106402.

[11] R. D. Liklikwatil, E. Noersasongko, and C. Supriyanto, "Optimasi K-Nearest Neighbor Dengan Particle Swarm Optimization Untuk Memprediksi Harga Komoditi Karet," *e-Jurnal JUSITI (Jurnal Sist. Inf. dan Teknol. Informasi)*, vol. 7–2, no. 2, pp. 172–182, 2018, doi: 10.36774/jusiti.v7i2.252.

[12] N. Bharanidharan and H. Rajaguru, "Classification of dementia using harmony search optimization technique," *IEEE Reg. 10 Humanit. Technol. Conf. R10-HTC*, vol. 2018-Decem, pp. 1–5, 2019, doi: 10.1109/R10-HTC.2018.8629846.

[13] L. S. Han and J. A. N. Nordin, "Predicting the stock price trends using a K-nearest neighbors-probabilistic model," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 18, pp. 6245–6255, 2018.

[14] M. E. Lasulika, "Prediksi Harga Komoditi Jagung Menggunakan K-Nn Dan Particle Swarm Optimazation Sebagai Fitur Seleksi," *Ilk. J. Ilm.*, vol. 9, no. 3, pp. 233–238, 2017, doi: 10.33096/ilkom.v9i3.148.233-238.

[15] B. Song, S. Tan, H. Shi, and B. Zhao, "Fault detection and diagnosis via standardized k nearest neighbor for multimode process," *J. Taiwan Inst. Chem. Eng.*, vol. 106, no. xxxx, pp. 1–8, 2020, doi: 10.1016/j.jtice.2019.09.017.

[16] A. Sharma, P. Madhushri, V. Kushvaha, and A. Kumar, "Prediction of the Fracture Toughness of Silicafilled Epoxy Composites using K-Nearest Neighbor (KNN) Method," *2020 Int. Conf. Comput. Perform. Eval. ComPE 2020*, pp. 194–198, 2020, doi: 10.1109/ComPE49325.2020.9200093.

[17] M. Bramer, *Principles of Data Mining*. Springer London, 2007.

e-ISSN: 2614-8404
p-ISSN:2776-3234

Combining Super Resolution Algorithm (Gaussian Denoising and Kernel Blurring) and Compare with Camera Super Resolution

Muhamad Ghofur[1*)], Tjong Wan Sen[2]
[1,2]Information Technology, Faculty of Computing, President University
Email: [1]muhamad.ghofur@student.president.ac.id, [2]wansen@president.ac.id

Abstract – This problem addresses the problem of low-resolution image (noisy) that will proof later by PSNR number. The best way to improve this low-resolution problem is by utilizing Super Resolution (SR) algorithm methodology. SR algorithm methodology refers to the process of obtaining higher-resolution images from several lower-resolution ones, that is resolution enhancement. The quality improvement is caused by fractional-pixel displacements between images. SR allows overcoming the limitations of the imaging system (resolving limit of the sensors) without the need for additional hardware. This research aims to find the best SR algorithm in form of stand-alone algorithm or combine algorithm by comparing with the latest SR algorithm (Camera SR) from the previous research made by Chang Chen et al in 2019. Furthermore, we confidence this research will become the future guideline for anyone who want to improve the limitation of their low-resolution camera or vision sensor by implementing those SR algorithms.

Keywords – Camera, Image, Resolution.

I. INTRODUCTION

Image resolution describes the amount of information contained by images. Lower resolution less would be the amount of information, higher resolution more would be amount of information in images. Resolution of a digital image can be classified in many ways: pixel resolution, spatial resolution, spectral resolution, temporal resolution, and radiometric resolution. In this resaerch concentration is given mainly in spatial resolution. Spatial resolution: A digital image is made from small picture elements called pixels. Spatial resolution refers to the pixel density in an image and measures in pixels per unit area.

The spatial resolution of an image is limited by the image sensors or the image acquisition devices. The modern image acquisition devices are using charge-coupled device (CCD) or complementary metal oxide semiconductor (CMOS) as active pixel sensor. These sensors are arranged in two dimensional arrays to capture two-dimensional image signals. The number of sensors per unit area or the sensor size determines the number of pixels in image. One way to increase the resolution of the imaging device is to increase the sensor density by reducing the size of sensors. When the size of sensors is reduced beyond a limit it causes shot noise in the captured images as reducing the size of sensor also reduces the amount of light incident on it. Increment in the number of sensors in imaging device/system also increases the hardware cost. Therefore, there is limitation with the hardware that restricts the spatial resolution of the image.

While spatial resolution is limited by sensor size, the image details (high frequency bands) are also limited by the optics due to lens blurs (associated with the sensor point spread function (PSF)), lens aberration effects, aperture diffractions and optical blurring due to motion. Constructing imaging chips and optical components to capture very high-resolution images is prohibitively expensive and not practical in most real applications, e.g., widely use surveillance cameras and cell phone built-in cameras. In some other scenarios such as satellite imagery, it is difficult to use high resolution sensors due to physical constraints.

Another way to address this problem is to accept the image degradations and use image processing to post process the captured images, to trade of computational cost with the hardware cost. These techniques are specifically referred as super resolution (SR) reconstruction.

Single image super-resolution (SR) is a typical inverse problem in computer vision. Generally, SR methods assume bicubic or Gaussian down sampling as the degradation model [1]. Based on this assumption, continuous progress has been achieved to restore a better high-resolution (HR) image from its low-resolution (LR) version, in terms of reconstruction accuracy [2],[3],[4],[5],[6],[7],[8],[9],[10],[11] or perceptual quality [12],[13],[14],[15],[16],[17],[18] However, these synthetic degradation models may deviate from the ones in realistic imaging systems, which results in a significant deterioration on the SR performance [19]. To better simulate the challenging real-world conditions, additional factors including noise, motion blur, and compression artifacts are integrated to characterize the LR images in either a synthetic [20] or a data-driven [21] manner.

II. RESEARCH METHODOLOGY

All methods that will be run in this research can be described in below process flow diagram.

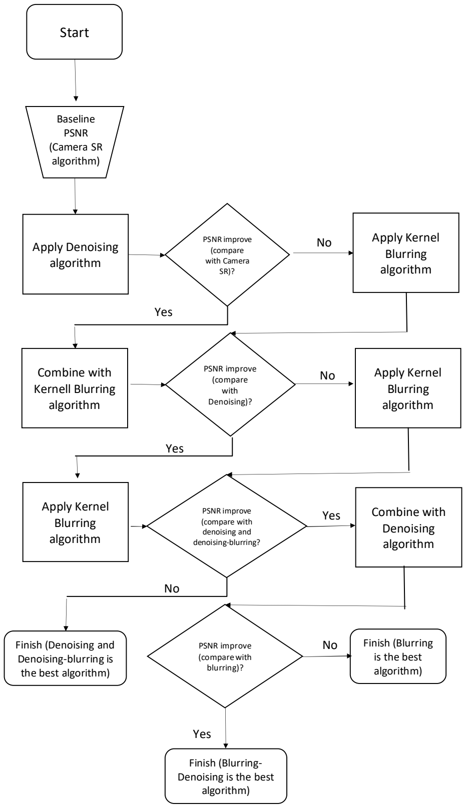

Start
Baseline PSNR (Camera SR algorithm)
Apply Denoising algorithm
PSNR improve (compare with Camera SR)?
No
Apply Kernel Blurring algorithm
Yes
Combine with Kernell Blurring algorithm
PSNR improve (compare with Denoising)?
No
Apply Kernel Blurring algorithm
Yes
Apply Kernel Blurring algorithm
PSNR improve (compare with denoising and denoising-blurring?
Yes
Combine with Denoising algorithm
No
Finish (Denoising and Denoising-blurring is the best algorithm)
PSNR improve (compare with blurring)?
No
Finish (Blurring is the best algorithm)
Yes
Finish (Blurring-Denoising is the best algorithm)

Figure. 1 Research Method.

All methods that will be run step by step as described in above process flow are explained below:

1. Get baseline picture sample.

In this method, we will use pictures from City100 data set. This is a set of pictures that were taken from some postcards and did it indoor. After that run Camera SR to these data set.

2. Apply SR Algorithm

After got the baseline data, then the next method that we will do is applying all those algorithms (stand-alone and combination). SR Algorithm that will applied in this research are Gaussian denoising, Kernel blurring, Denoising-Blurring and Blurring-Denoising.

After that, every time we finished to apply each algorithm, we will compare their PSNR score with Camera SR PSNR score to get which one is the best SR algorithm if we refer with their PSNR score.

III. Results And Discussion

We will start with showing some pictures from City100 data set and our CCD picture dataset. This will become base line picture data. Then after that it will proceed with Super Resolution algorithm in sequence and vice versa. We have set a requirement in this research method, that if there is any algorithm which does not give improvement in the picture data set then we will continue with this algorithm.

Figure. 2 Five test images from city 100 data set.

A. Kernel Blurring

A kernel is a (usually) small matrix of numbers that is used in image convolutions, while convolution itself is a general-purpose filter effect for images. Convolution filtering is usually used to modify the spatial frequency characteristics of an image. They did it by applying a matrix to an image and a mathematical operation comprised of integers. The mathematical operation will work by determining the value of a central pixel by adding the weighted values of all its neighbors together. Then the output is a new modified filtered image. Different sized kernels containing different patterns of numbers produce different results under convolution. The image blurring process is commonly modeled as the convolution of a clear image with a shift-invariant kernel plus noise, i.e.,

$$f = k * g + n \tag{1}$$

where $*$ denotes the discrete convolution operator, g denotes the clear image, g denotes the available blurry observation, k denotes the blur kernel, and n denotes the image noise. The size of a kernel is arbitrary but 3X3 is often used.

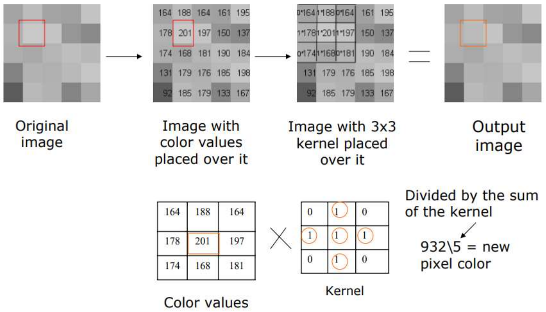


Figure. 3 Kernel matrix.

Based on our research, to get higher PSNR we must use a square kernel matrix (3X3, 5X5, etc.). When we did not use a square kernel matrix, the color matrix distribution will not even and created some pictures spot greyish or darker. Based on our research, the best kernel matrix square that get the highest PSNR is 5X5 matrix.

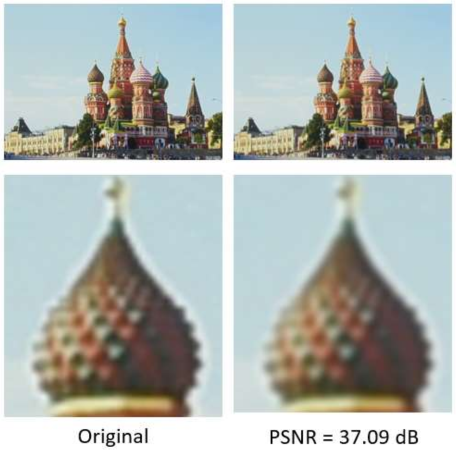


Figure. 4 Kernel blurring PSNR result.

B. Gaussian Denoising

The probability density function p of a Gaussian random z variable is given by:

$$pG(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (2)$$

where z represents the grey level, μ the mean grey value and σ its standard deviation. In digital image processing Gaussian noise can be reduced using a spatial filter, though when smoothing an image, an undesirable outcome may result in the blurring of fine-scaled image edges and details because they also correspond to blocked high frequencies. The different with kernel blurring is in denoising the result will be based on how big the noise filter. The bigger noise filter, the bigger high frequency cannot be passed. Based on our research the noise filter = 25 gave the highest PSNR result.

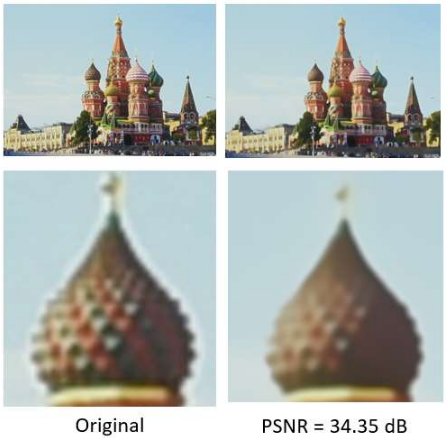


Figure. 5 Gaussian denoising PSNR result.

C. Blurring and Denoising

Here are the PSNR results after we combine Blurring and Denoising.

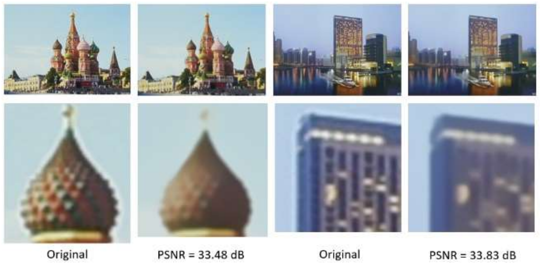


Figure. 6 Blurring and denoising PSNR result.

D. Denoising and Blurring

Here are the PSNR results after we combine Denoising and Blurring.

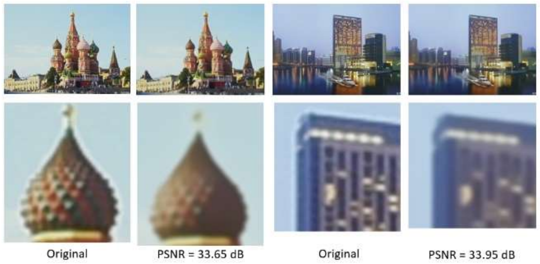


Figure. 7 Denoising and blurring PSNR result.

E. Result Summary

After we include with the result from previous research by using Camera SR methodology, here is the PSNR result summary from each algorithm methodology and their combination.

Table 1. Research Summary

Picture	Blurring	Denoising	Blurring - Denoising	Denoising - Blurring	Camera SR
St. Petersburg.	37.09	34.35	33.48	33.65	31.00
Dubai	38.77	34.50	33.83	33.95	31.94
Venice	35.00	31.46	30.52	30.68	28.19
Rome	37.85	34.42	33.46	33.64	33.04
New York	34.96	32.64	31.50	31.68	27.14
Average	36.73	33.47	32.56	32.72	30.26

By looking at the PSNR (dB) result in above table we can see that by our research, with picture characteristics in City100 data set, by only utilize one algorithm methodology that is Kernel Blurring actually we can get the best picture quality. How it can be happened? Because theoretically, a kernel blurring is a (usually) small matrix of numbers that is used in image convolutions, while convolution itself is a general-purpose filter effect for images. By using Kernel, a single matrix color can distribute some parts of their color to their neighborhood matrix. It is not just like a denoising algorithm when they want to be smoothing the noise using their spatial filter, an undesirable outcome may result in the blurring of fine-scaled image edges and details because they also correspond to blocked high frequencies (due to their power of noise filter).

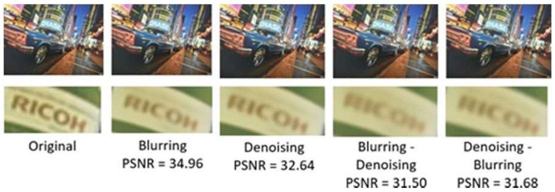


Figure. 8 SR algorithm and their combination comparison.

However, for all SR algorithm and combination if we look at the pictures result each of them, we can see that some effect may happened when we zoom it in a specific point. We can get a blurred fine scaled image even though in overall picture result we can see a better-quality picture. This is the important thing we may consider when we want to implement these Super Resolution algorithm for our image processing. Because in some cases when you already have a clean (free or less noise) picture, you may not need these SR algorithms anymore. Or perhaps you need another SR algorithm like Interpolation for the example if you want to play with bigger pixel picture.

IV. Conclusion

In this research we try to analyze the benefit of some Super Algorithm methodology and their combination to find solution for pictures with noise characteristics. We also want to compare these solutions with Camera SR methodology that previously has been proposed by Chen et al. We use the same data set that is City100 that been developed by Chen et al in their paper in 2019. Based on City100, we analyze the advantage and disadvantage of some SR algorithm including Camera SR methodology and validate that actually one of the algorithms that is Kernel Blurring is the best algorithm of other existing SR methods with average PSNR 36.73 dB and a practical solution so far to boost the performance of noisy pictures. Perhaps Chen et al. did not consider "Kernel Blurring" algorithm in their paper as we did not see this algorithm was discussed and analyzed in their paper. Despite the validating result, there are still some updated methodology and picture aspects that we have not touched in our research. Like for example how to increase pixel quantity, because increasing pixel quantity is also part of Super Resolution works. While in the updated methodology, perhaps we also can utilize the use of Deep Learning methodology to predict how Neural Network in a computer can do an automatic convolutional filter. The above example will be considered as our future works.

References

[1] K. Zhang, W. Zuo, and L. Zhang, "Learning a Single Convolutional Super-Resolution Network for Multiple Degradations," 2018, doi: 10.1109/CVPR.2018.00344.

[2] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep Back-Projection Networks for Super-Resolution," 2018, doi: 10.1109/CVPR.2018.00179.

[3] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-December, doi: 10.1109/CVPR.2016.182.

[4] W. S. Lai, J. Bin Huang, N. Ahuja, and M. H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-January, doi: 10.1109/CVPR.2017.618.

[5] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017, vol. 2017-July, doi: 10.1109/CVPRW.2017.151.

[6] A. Shocher, N. Cohen, and M. Irani, "Zero-Shot Super-Resolution Using Deep Internal Learning," 2018, doi: 10.1109/CVPR.2018.00329.

[7] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-January, doi: 10.1109/CVPR.2017.298.

[8] T. Tong, G. Li, X. Liu, and Q. Gao, "Image Super-Resolution Using Dense Skip Connections," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-October, doi: 10.1109/ICCV.2017.514.

[9] Z. Xiong, X. Sun, and F. Wu, "Robust web image/video super-resolution," *IEEE Trans. Image Process.*, vol. 19, no. 8, 2010, doi: 10.1109/TIP.2010.2045707.

[10] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Lecture*

Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 11211 LNCS, doi: 10.1007/978-3-030-01234-2_18.

[11] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual Dense Network for Image Super-Resolution," 2018, doi: 10.1109/CVPR.2018.00262.

[12] A. Bulat and G. Tzimiropoulos, "Super-FAN: Integrated Facial Landmark Localization and Super-Resolution of Real-World Low Resolution Faces in Arbitrary Poses with GANs," 2018, doi: 10.1109/CVPR.2018.00019.

[13] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11133 LNCS, doi: 10.1007/978-3-030-11021-5_21.

[14] X. Deng, "Enhancing Image Quality via Style Transfer for Single Image Super-Resolution," *IEEE Signal Process. Lett.*, vol. 25, no. 4, 2018, doi: 10.1109/LSP.2018.2805809.

[15] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9906 LNCS, doi: 10.1007/978-3-319-46475-6_43.

[16] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-January, doi: 10.1109/CVPR.2017.19.

[17] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch, "EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-October, doi: 10.1109/ICCV.2017.481.

[18] X. Wang, K. Yu, C. Dong, and C. Change Loy, "Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform," 2018, doi: 10.1109/CVPR.2018.00070.

[19] T. Michaeli and M. Irani, "Nonparametric blind super-resolution," 2013, doi: 10.1109/ICCV.2013.121.

[20] R. Timofte, S. Gu, L. Van Gool, L. Zhang, and M. H. Yang, "NTIRE 2018 challenge on single image super-resolution: Methods and results," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2018, vol. 2018-June, doi: 10.1109/CVPRW.2018.00130.

[21] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a GAN to learn how to do image degradation first," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11210 LNCS, doi: 10.1007/978-3-030-01231-1_12.

e-ISSN: 2614-8404
p-ISSN : 2776-3234

The Prediction of Gold Price Movement by Comparing Naive Bayes, Support Vector Machine, and K-NN

Yahya Suryana[1*)], Tjong Wan Sen[2]
[1,2]informatics Engineering Study Program, Faculty of Computer, President University Email:
[1]yahya.surryana@student.president.ac.id, [2]wansen@president.ac.id

Abstract – Gold is a yellow precious metal that can be forged so it is easy to form with various forms of jewelry such as pendants, earrings, rings, bracelets and others, gold has a high value. Gold itself is an exchange rate used in ancient times before the existence of money as it is today. Gold also can be used as an investment that is profitable for the investor and it has less risks. Investment is a form of fund management to give benefit by putting fund in allocation that is predicted will give additional benetifs. Prediction of gold price movements or predictions of gold price in gold stock investment, this research uses 3 (three) algorithms that will be implemented in analysis and increase accuracy, in the discussion or research that was made using the Naïve Bayes algorithm, Support Vector Machine and K-Nearest Neighbor, the dataset is obtained from the website, namely www.finance.yahoo.com the data was then tested using Rapid miner tools so that the average value of the Support Vector Machine algorithm with an accuracy rate of 57.59%, precision 58 ,73% and recall 51,78%. The next is the Naïve Bayes algorithm so that it is known to have an accuracy rate of 55.59%, precision 54.55% and recall 51.70%. Based on the comparison of the three algorithms, it is known that the one with the best accuracy, precision, and recall is the K-NN algorithm with 61.90% accuracy, 60.98% precision, and 60.35% recall. Furthermore, the results of testing the K-Nearst Neighbor algorithm have good results compared to the 3 (three) other algorithm tests and the Naïve Bayes algorithm testing has a low level of accuracy, namely 55.59%, precision 54.55% and recall 51.70%. The research uses 3 algorithms, namely naive bayes, K-nearst neighbor and Support Vector Machine, because the three algorithms are well-established algorithms to be applied to research, especially in time series gold price research and are very good, especially for classification.

***Keywords** – Data Mining, Naïve Bayes, KNN. Support Vector Machine.*

I. INTRODUCTION

Gold is one of the most malleable and highly malleable yellow precious metals. Gold is the exchange rate used before the existence of money as it is today therefore the risk impact of investing in gold is very small, this yellow precious metal has two types, namely, gold for investment only, gold for jewelry only such as necklaces, rings, bracelets, earrings and others, gold is investment in gold stocks or gold futures. Investment process contains risk and uncertainty. The investment that everyone can do is a gold investment, so this so that gold becomes an investment that is in great demand and becomes a prima donna among the upper class, upper or lower middle, but basically in investing in gold this is when price fluctuations occur every day, every month or even every year, this fluctuation risk is called time series risk, where the price is always going up and down. In order not to happen and avoid risk, in this study a strong prediction is made about the price of gold by using time series data so that gold investors can know when to invest and when to resell so that they can provide benefits for gold investors according to the plans that have been made, This gold price forecast is made so that investors get profits and according to the plans made in this research. This gold precious metal investment is more profitable, However, investors must know that there are several factors that influence when the gold exchange rate occurs Gold is one of the most valuable commodities and is traded in various parts of the world, such as one of the leading countries, namely Saudi Arabia, with the most investors reaching 70%. Gold is synonymous with a symbol of luxury and glory, but gold is vulnerable and heavily influenced by various economic indicators such as interest rate, inflation, and Bruto Domestic Product (Produk Domestik Bruto). Certain information is utilized its knowledge. One of the approach that can be used to analize a group. Several research methods have been carried out to predict gold prices, one of which is the Prediction of Gold Price Movements in Gold Stock Investments By Comparing the Naïve Bayes Algorithm, KNN and Support Vector Machine[1]. The Naïve Bayes algorithm has been used to predict gold prices as was done by Mohammad Guntur, Yulius Santony and Yuhandri in 2018 with the title Gold Price Prediction Using The Naïve Bayes Method In Investing To Minimize Risk[2] with the results of his research, namely predicting the price of gold which can help make decisions in determining whether to sell or give gold to predict the price of gold for the next 14 days The data used for testing are 16 data and an accuracy of 75% [3] is obtained, but Naïve Bayes has many deficiencies in classification problems, classification so that further research is needed to improve accuracy. In this study, several tests were carried out on the price of gold by adding a dataset so as to get high accuracy Based on the brief discussion of the above problems, predictions are made by increasing accuracy of the gold price prediction, This research uses gold price data with of gold time series data obtained from a time series gold data website, namely www.finance.yahoo.com and processing gold data so that the results of gold price predictions can be known. Based on the explanation described above, this research was conducted so that compare Naïve Bayes[4], Support Vector Machine[5], and K-Nearest Neighbor algorithm[6],[8],[9],[10]".

II. RESEARCH METHODOLOGY

In this study, analysis and methodology were carried out on the gold time series datafacilites research and able to run systemmatically fulfill the purpose as expected therefore steps in research research stage is made and will be run as follows

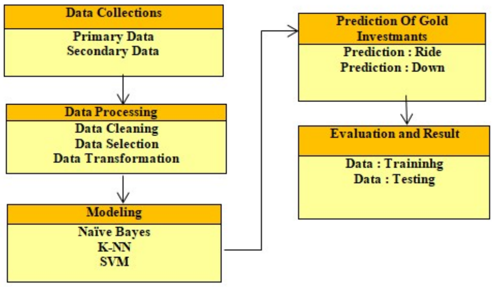


Figure 1 Research Stages

2.1. Data collection

At the stage of data collection, research is carried out using 2 (two) references to the types of data obtained, including:

1. Primary Data

The data taken in this study is data from a website www.finance.yohoo.com, namely the price of gold time series for 5 (five) years, from 2014 to 2019, this data is used as study material for learning and training in research. next.

At this stage, namely taking existing references, namely looking for references in books or journals and also on papers that are similar to the research being conducted.

2.2 Data Processing

At this stage, data processing is carried out by pre-processing the data to be processed. The data obtained in this study are time series gold price data taken from a website www.finance.yahoo.com as an initial stage, the processing is carried out in 3 stages, the first stage is the elimination of some noise data, then the second stage namely dividing into training data and testing data, 90% for training and 10% for testing data. The third stage of processing is carried out using rapidminer tools to find accuracy, precision and recall on rapidminer tools.

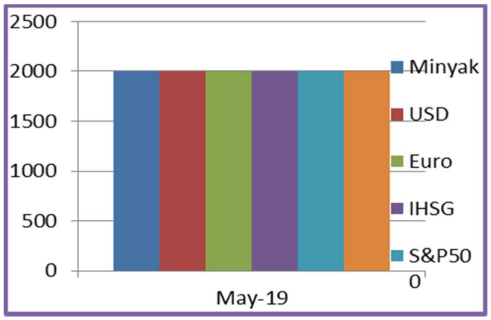


Figure 2 Dataset Graph

At the data cleaning stage, it aims to clean up inconsistent data empty values or commonly called empty tuples, duplicate data and correct data errors, the data repair process is carried out manually, with the help of spreadsheet tools

2.4 Data Selection

Data selection is the process of selecting data from existing operational data before entering the data and information mining stage. At this stage, the following steps will be carried out:

1. The data sample is taken randomly with the attribute parameters on the data www.finance.yahoo.com. The document that has the largest amount of data to be used as a dataset and ensures that the selected data is suitable for use in the modeling process.
2. After viewing this dataset, you will get the amount of testing data.
3. Choose the attributes that will be used and analyzed, because in the initial data there are some unneeded attributes such as attributes.

Table 1 Data and Attribute Used

No	**Atribute**						
	Date	***Oil***	***USD***	***Euro***	***IHSG***	***S&P500***	***Gold***
1	31/05/19	53.5	14385	15,964.	6209	2,752.0	1,30
2	30/05/19	56.5	14385	16,027	6209	2,788.8	1,28
3	14/05/10	58.8	14478	16,125	5907	2,840.2	1,27

2.5 Transformation Data

The data transformation stage is the process of changing the initial data format into a standard data format for the process of reading data using the algorithms in the programs or tools used.

2.6 Modelling

The modeling in this study was carried out using data mining classification techniques for the Naïve Bayes algorithm, Support Vector Machine, and K-Nearest Neighbor. This technique was chosen because it is a commonly used method in data mining research to classify or recognize new data that has never been studied, especially in predicting gold price movements or predicting gold prices in gold stock investments. The The algorithm that will be used for the analysis of this research is the Naïve Bayes algorithm, Support Vector Machine, and K-Nearest Neighbor [11],[12],[13],14]. These 3 (three) algorithms are algorithms that have been established and is widely implemented in classification techniques. In addition, this algorithm has advantages, namely in the form of good accuracy in handling a processed dataset

.

2.7 Testing and Validation of Reseach

Method testing is carried out with the aim of knowing the results of the analyzed calculations and measuring the methods and algorithms used whether they function properly or not. The testing process uses the rapidmener tool and sees whether the data is in accordance with the results obtained through the tool. While the validation of the methods and algorithms of Naïve Bayes[15],[16],[17], Support Vector Machine[18],[19],[20], and K-Nearest Neighbor[21],[22] is done by measuring the results of accuracy, percision and recall and can be calculated using the Confusion Matrix as follows:

2.8 Proposed Method

The method of this research uses 3 (three) algorithms, namely K-NN, Naive Bayes and Support Vector Machine to train data accuracy in time series gold price prediction assisted by rapid miner software.

2.9 K-fold Cross Validation

Using K-Fold Cross-Validation for statistical analysis that will be generalized to independent data sets (Suyanto, 2019). This technique is mainly used to make model predictions and estimate how accurate a predictive model will be when run in practice. The purpose of defining to test the model phase is, data validation, to limit problems such as the occurrence of such To provide insight into how the model will generalize independently of the dataset (i.e., unknown dataset, for example from a real problem), the following is an example of a data iteration table on K-Fold Cross Validation.

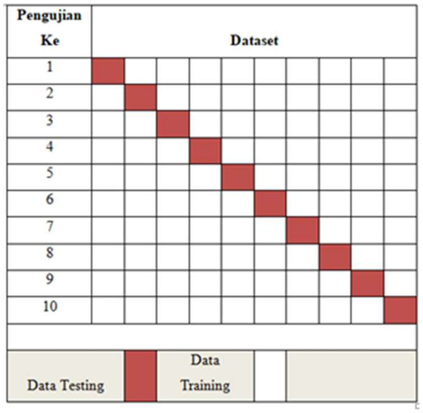


Figure 3 Testing With 10 K-fold Cross Validation

1. KNN

The accuracy value is calculated by adding up the correct data, namely positive (True Positive) plus true negative (True Negative) divided by the number of correct data, namely true positive (True Positive), true negative (True Negative) and added by the false data which is positive (False Positive), Negative (False Negative).

Table 2 Confusion Matrix (K-NN) Accuracy Calculation Formula

Score Prediction	**True Value**	
	Calculation K-NN	
	True	***False***
True	TP 11	FP 3
False	FN 0	TN 0

$$\text{Accuracy } \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} 100 \tag{1}$$

$$= \frac{11 + 0}{11 + 0 + 3 + 0} * 100\%$$

$$= \frac{11}{14} * 100\%$$

$$= 0{,}7857142857 * 100\%$$

$$= 78{,}57\%$$

The precision value is calculated by dividing the

number of true positive data (True Positive) divided by the number of true positive data (True Positive) and false positive data (False Positive).

$$\text{Precision } = \frac{\text{TP}}{\text{TP} + \text{FP}} 100\% \tag{2}$$

$$= \frac{11}{11 + 3} * 100\%$$

$$= \frac{11}{14} * 100\%$$

$$= 0{,}7857142857 * 100\%$$

$$= 78{,}57\%$$

The recall value is calculated by dividing the correct

data, which is true positive (True Positive) and divided by the number of correct data, namely true positive (True Positive) and false data, namely false negative (False Negative).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100\% \tag{3}$$

$$= \frac{11}{11 + 0} * 100\%$$

$$= \frac{11}{11} * 100\%$$

$$= 1 * 100\%$$

e-ISSN: 2614-8404
p-ISSN : 2776-3234

$$= 100\%$$

The accuracy value is calculated by adding up the correct data, namely true positive (True Positive) plus the true negative value (True Negative) also divided by the number of correct data, namely true positive (True Positive), true negative (True Negative) and added to the data False is false positive (False Positive), and the data is divided by false negative (False Negative).

Tabel 3 Confusion Matrix SVM Precision Calculation SVM

$$\mathrm{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} * 100\% \qquad (4)$$

$$= \frac{11+0}{11+0+3+0} * 100\%$$

$$= \frac{11}{14} * 100\%$$

$$= 0{,}7857142857 * 100\%$$

$$= 78{,}57\%$$

The precision value is calculated by dividing the number of true positive data (True Positive) divided by the number of true positive data (True Positive) and false positive data (False Positive)

$$.\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} * 100\% \qquad (5)$$

$$= \frac{11}{11+3} * 100\%$$

$$= \frac{11}{14} * 100\%$$

$$= 0{,}7857142857 * 100\%$$

$$= 78{,}57\%$$

The recall value will be calculated by dividing the correct data, namely true positive (True Positive) then dividing by the correct number of data, namely true positive (True Positive) and incorrect data, namely false negative (False Negative).

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} * 100\% \qquad (6)$$

$$= \frac{11}{11+0} * 100\%$$

$$= \frac{11}{11} * 100\%$$

$$= 1 * 100\%$$

$$= 100\%$$

The accuracy value is calculated by adding up the correct data which is positive (True Positive) plus the negative value (True Negative) divided by the number of correct data which is positive (True Positive), Negative (True Negative) and added by false data which is positive (False Positive), Negative (False Negative).

Table 4 Confusion Matrix Accuracy Calculation Naïve bayes

Score Prediction	**True Value**	
	Calculation Naïve Bayes	
	True	***False***
True	TP 7	FP 3
False	FN 4	TN 0

Score Prediction	**True Value**	
	Calculation SVM	
	True	***False***
True	TP 7	FP 3
False	FN 4	TN 0

$$\mathrm{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} * 100\% \qquad (7)$$

$$= \frac{7+0}{7+0+3+4} * 100\%$$

$$= \frac{7}{14} * 100\%$$

$$= 0{,}5 * 100\%$$

$$= 50\%$$

The precision value is calculated by dividing the number of true positive data (True Positive) divided by the number of true positive data (True Positive) and false positive data (False Positive).

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} * 100\% \qquad (8)$$

$$= \frac{7}{7+3} * 100\%$$

$$= \frac{7}{10} * 100\%$$

$$= 0{,}7 * 100\%$$

$$= 70\%$$

The recall value is calculated by dividing the correct data which is positive (True Positive) with the sum of the correct data which is positive (True Positive) and the incorrect data which is negative (False)..

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100\% \tag{9}$$

$$= \frac{7}{7+4} * 100\%$$

$$= \frac{7}{11} * 100\%$$

$$= 0{,}636363664 * 100\%$$

$$= 63.64\%$$

From the sample data as many as 14 data, then the results from these data state the level of accuracy, recall and persicion of the K-NN, SVM and Naïve Bayes algorithms. The following are the results of the accuracy, recall and persicion values

Table 5 Results of accuracy, recall and precision values

No	Result			
	Algorithm	*Accuracy*	*Precision*	*Recall*
1	K-NN	78.57%	78.57%	100%
2	SVM	78.57%	78.57%	100%
3	Naïve Bayes	50%	70%	63.64%

III. Results And Discussion

In this study using the SVM, Naïve Bayes, K-NN algorithm which will be tested so that it will get the results of accuracy, precision, and recall values as well as predictions that can be used in making decisions when investing in gold. The source of data as an object in this study is gold price data over a period of years The data used in this study consists of attributes or variables such as, Oil, USD, Euro, IHSG, S&P500, Gold, so that it can find out a result of knowledge in predicting gold data and in this test using 2000 data and the data is divided into two 90% training data and 10% testing data, so that it can produce models or values of accuracy, precision and recall obtained from the three algorithms to test gold data, the data used is divided randomly or randomly into subsets, namely t1, t2, t3, ... , t10 with the same data size.

From the 200 testing data and 1800 training data, data modeling will be formed to be tested by rapidminer in finding results so that they can find out the accuracy, precision and recall results. and compare the 3 algorithms that have good results or have high levels of accuracy, precision and recall

from the test results can produce information and knowledge in predicting gold investment. The following is Table 6 of the modeling process in testing gold data.

Tabel 6. 10 x test

No	Testing					
	10 x test					
Test 1	200	200	200	200	200	200
Test 2	200	200	200	200	200	200
Test 3	200	200	200	200	200	200
Test 4	200	200	200	200	200	200
Test 5	200	200	200	200	200	200
Test 6	200	200	200	200	200	200
Test 7	200	200	200	200	200	200
Test 8	200	200	200	200	200	200
Test 9	200	200	200	200	200	200
Test 10	200	200	200	200	200	200

Based on table 7 it is shown that the p value used by the testing process. Here are the steps for testing the data with 10 tests.

Table 7 Sharing of Training Data and Testing Data

No	Data Training and Data Testing	
	Data Training	*Data Testing*
Test 1	1800	200
Test 2	1800	200
Test 3	1800	200
Test 4	1800	200
Test 5	1800	200
Test 6	1800	200
Test 7	1800	200
Test 8	1800	200
Test 9	1800	200
Test 10	1800	200

The data used is divided into 10 parts, namely p1, p2,….p10 and training data, = (1, 2, 3, 4, 5, 6, 7……, 10) is used as testing data and others as training data.

Accuracy results are calculated for each test (test-1, test-2, test-3, test-4, test-5, test-6, test-7, test-8, test-9, test-10) then calculate the average level of accuracy from all tests to get the level of accuracy, precision and recall of the overall data

The data is saved in excel workbook format which is then converted into a data frame with the read excel command.

Table 8 Sample Data Testing

No	Atribute						
	Date	*Oil*	*USD*	*Euro*	*IHSG*	*S&P500*	*Gold*
1	31/05/19	53.5	14385	15,964.	6209	2,752.0	1,30
2	30/05/19	56.5	14385	16,027	6209	2,788.8	1,28
3	29/05/19	591	144	16,028	6104	2,783.0	1,28
4	28/05/19	591	143	16,043	6033	2,808.3	1,27
5	27/05/19	591	143	16,103	6098	2,802.1	1,28
6	26/05/19	587	143	16,129	6098	2,808.3	1,28
7	25/05/19	586	143	16,126	6098	2,802.1	1,28
8	24/05/19	586	144	16,126	6057	2,826.0	1,28
9	23/05/19	579	145	16,167	6032	2,822.2	1,28
10	22/05/19	614	144	16,199	5939	2,856.2	1,27
11	21/05/19	629	144	16,160	5951	2,864.3	1,27
12	20/05/19	631	144	16,147	5907	2,8402	1,27
13	19/05/19	627	144	16,129	5907	2,840.2	1,27
14	14/05/10	58.8	14478	16,125	5907	2,840.2	1,27

The training and testing data will be processed using the SVM, Naïve Bayes and K-NN algorithms, the data is tested with the rapidminer tools, this study will evaluate the classification results of the data, the real data that have been tested, so that the results of accuracy, recall and precision can be seen on tools rapidminer determining data result of powder coating production. Here is the overall process of data testing by using rapid minerr. From the testing result with 10 testings randomly has generates highest accuracy level, it is K-NN algorithm.

Table 9 Accuraci Value Result

No	SVM		Naïve Bayes		K-NN	
	Accuracy		*Accuracy*		*Accuracy*	
1	Test 1	58.50%	Test 1	60.00%	Test 1	63.00%
2	Test 2	59.50%	Test 2	60.00%	Test 2	63.00%
3	Test 3	60.50%	Test 3	59.50%	Test 3	58.50%
4	Test 4	53.50%	Test 4	48.50%	Test 4	62.00%
5	Test 5	55.50%	Test 5	57.00%	Test 5	59.00%
6	Test 6	54.50%	Test 6	51.00%	Test 6	66.00%
7	Test 7	55.00%	Test 7	55.50%	Test 7	62.00%
8	Test 8	52.00%	Test 8	43.50%	Test 8	53.50%
9	Test 9	64.50%	Test 9	55.00%	Test 9	62.00%
10	Test 10	66.00%	Test 10	67.50%	Test 10	70.00%

From the test results by conducting 10 random tests,the highest accuracy level is the K-NN . algorithm

Table 10 Precision Value Result

No	Algorithm					
	SVM		*Naïve Bayes*		*KNN*	
1	Precision		Precision		Precision	
2	T 1	61.40%	T 1	66.67%	T 1	66.67%
3	T 2	59.72%	T 2	55.68%	T 2	55.68%
4	T 3	55.30%	T 3	64.76%	T 3	64.76%
5	T 4	53.66%	T 4	58.89%	T 4	58.89%
6	T 5	54.95%	T 5	66.33%	T 5	66.33%
7	T 6	52.63%	T 6	59.09%	T 6	59.09%
8	T 7	56.00%	T 7	49.41%	T 7	49.41%
9	T 8	68.52%	T 8	58.62%	T 8	58.62%
10	T 9	62.50%	T 9	65.96%	T 9	65.96%
11	T 10	62.50%	T 10	65.96%	T 10	65.96%

From the testing result with 10 testings randomly has generates highest precision level, it is K-NN algorithm.

Table 8 Recall Value Result

No	SVM		NB		KNN	
	Recall		*Recall*		*Recall*	
1	T 1	50.00%	T 1	51.92%	T 1	64.42%
2	T 2	87.50%	T 2	74.17%	T 2	76.67%
3	T 3	46.24%	T 3	51.61%	T 3	52.69%
4	T 4	68.22%	T 4	53.27%	T 4	63.55%
5	T 5	67.35%	T 5	62.24%	T 5	54.08%
6	T 6	50.00%	T 6	30.00%	T 6	65.00%
7	T 7	21.74%	T 7	79.35%	T 7	56.52%
8	T 8	25.00%	T 8	23.91%	T 8	45.65%
9	T 9	40.66%	T 9	29.67%	T 9	56.04%
10	T 10	61.11%	T 10	61.11%	T 10	68.89%

From the testing result with 10 testings randomly has generates highest recall level, it is K-NN algorithm. According to the data that has been tested, the result of the data explains the level of accuracy, recall, and precision. Here below is the graphic of overall data that has been tested by using rapid miner.

1. Accuracy

The graphic result of accuracy value with result of data training and testing.

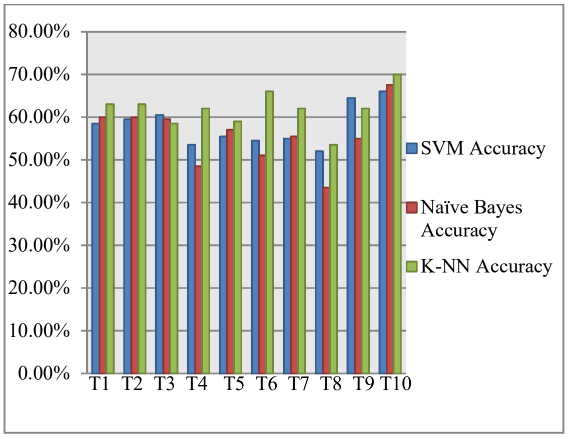


Figure 4 Grafik Accuracy

2. Precission

The graphic result of precision value with result of data training and testing.

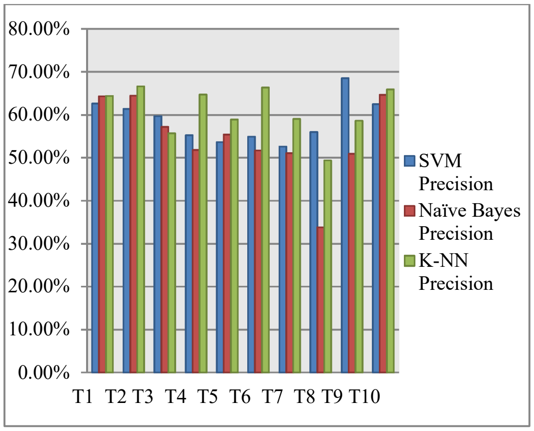


Figure 5 Grafik Precision

The graphic result of recall value with result of data training and testing.

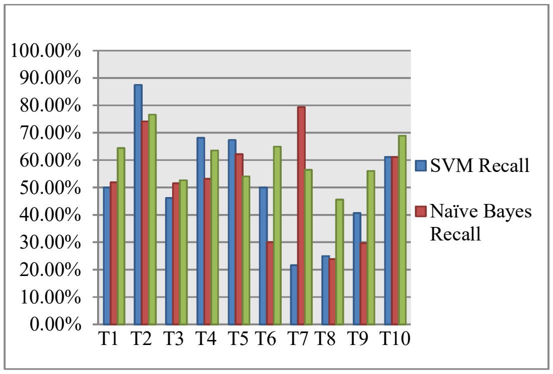


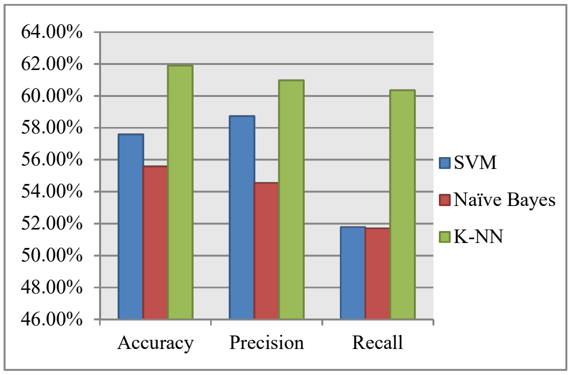
Figure 6 Grafik Recall

The test results from the gold data produce the level of accuracy, precision and recall of each algorithm, which can be explained as follows. The SVM algorithm has an accuracy rate of 57.59%, precision 58.73% and recal 51.78% while the Naïve Bayes algorithm has an accuracy rate of 55.59% precision 54.55% and recall 51.70% and which has a level accuracy, The best precision and recall in comparing the 3 algorithms in testing this gold data is the K-NN algorithm which has an accuracy value of 61.90%, precision 60.98% and recall 60.35% and can be seen from table 4.7 the results of 3 the algorithm.

Table 8 Accuracy Value Resilt

No	Accuracy			
	Algorithm	*Accuracy*	*Precissin*	*Recall*
1	Naïve Bayes	55.59%	54.55%	51.70%
2	SVM	57.59%	58.73%	51.78%
3	KNN	61.90%	60.98%	60.35%

From the test results by conducting 10 random tests with the highest level of accuracy, precision and recall being K-NN, the following is an overview of the graph of the results of testing 3 algorithms.



Picture 6 Graphic of Testing Result Value

Analysis

Based on the results of the gold data analysis and the results of the gold data testing that was carried out, the data testing resulted in good results with the composition of the 6 data attributes tested, then the data was tested with the SVM algorithm, K-NN and Naïve Bayes produce good results above 50% and based on the results obtained in this study, the algorithm that produces good accuracy, precision and recall levels of the 3 algorithms is K-NN. In gold investment so that we can analyze and the results obtained from testing by doing random or random tests by generating the average value of the algorithm SVM has 57.59% accuracy, 58.73% precision and 51.78% recall, while the Naïve Bayes algorithm has 55.59% accuracy, 54.55% precision and 51.70% recall and which has the best level of accuracy, precision and recall in comparing the 3 algorithms in testing this gold data is the K-NN algorithm which has 61.90% accuracy, 60.98% precision and 60.35% recall. The results of the K-NN algorithm have quite good results from the 3 tests of the algorithm.

From the testing of the nave Bayes algorithm, it was carried out randomly or randomly in the test and the results of the nave Bayes algorithm test had an accuracy value of 55.59%, precision 54.55% and recall 51.70%. From the prediction results, the data tested or read from the Naive Bayes algorithm has a fairly low gold data classification value or in the process of predicting data up will be read down and data down will be read up, so from the results of this study, random testing was carried out with the same data using 200 testing data and had 6 attributes used because gold data in general can be categorized as good results because the data has 6 attributes that have information about gold investment.

IV. Conclusion

Based on the test results using the SVM, K-NN and Naïve Bayes algorithms, the following conclusions can be drawn:

In testing the gold data, the data tested resulted in good results with the composition of the 6 data attributes being tested, then the data was tested with the SVM, K-NN and Naïve Bayes algorithms produce good results above 50%, testing by conducting random tests or random by producing an average value of the SVM algorithm has an accuracy rate of 57.59%, precision 58.73% and recall 51.78% while the Naïve Bayes algorithm has an accuracy level of 55.59%, precision is 54.55% and recall is 51.70% and which has a level of accuracy, The best precision and recall in comparing the 3 algorithms in testing this gold data is the K-NN algorithm which has 61.90% accuracy, 60.98% precision and 60.35% recall. the results of the K-NN algorithm have fairly good results from the 3 tests of the algorithm and the naesve Bayes algorithm test has an accuracy value of 55.59%, precision 54.55% and recall 51.70% From the prediction results, the data tested or read from the naïve Bayes algorithm has a fairly low gold data classification value or in the process of predicting the data up will be read down and the data down will be read up. Based on the research conducted, this research can provide some suggestions as follows:

Maximize or add more specific and more attributes in Naïve Bayes, SVM and K-NN classifications, further research is needed by testing with other algorithms such as C.45, C.50 and so on in order to obtain comparisons with the highest level of accuracy in making classifications on Naïve Bayes, SVM and K-NN and further research to improve accuracy, precision and recall in classifying by conducting experiments on each parameter.

REFERENCES

[1] Guntur, M., Santony, J., & Yuhandri, Y. (2018). Prediksi Harga Emas dengan Menggunakan Metode Naïve Bayes dalam Investasi untuk Meminimalisasi Resiko. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *2*(1), 354-360

[2] Mahena, Y., Rusli, M., & Winarso, E. (2015). Prediksi Harga Emas Dunia Sebagai Pendukung Keputusan Investasi Saham Emas Menggunakan Teknik Data Mining. *Kalbiscentia J. Sains dan Teknol*, *2*(1), 36-51.

[3] Saputro, N. D. (2015). Penerapan Algoritma Support Vector Machine untuk Prediksi Harga Emas. *Jurnal Informatika Upgris*, *1*(1 Juni).

[4] Witjaksono, A. A. (2010). *Analisis Pengaruh Tingkat Suku Bunga SBI, Harga Minyak Dunia, Harga Emas Dunia, Kurs Rupiah, Indeks Nikkei 225, dan Indeks Dow Jones terhadap IHSG (studi kasus pada IHSG di BEI selama periode 2000-2009)* (Doctoral dissertation, UNIVERSITAS DIPONEGORO)

[5] Budiyono, E. P., Nerfita Nikentari, S. T., Sallu, S., & Kom, S. ANALISA KLASIFIKASI KADAR KARAT EMAS MENGGUNAKAN METODE K-NEAREST NEIGHBOURS (KNN).

[6] Hidayat, R. N. (2013). *Implementasi Jaringan Syaraf Tiruan Perambatan Balik untuk Memprediksi Harga Logam Mulia Emas Menggunakan Algoritma Lavenberg Marquardt* (Doctoral dissertation, Diponegoro University).

[7] Fajrul Falah (2015). Rancang bangun aplikasi prediksi pergerakan harga emas logam mulia dengan menggunakan metode Backpropagation

[8] Sari, Y. (2017). Prediksi Harga Emas Menggunakan Metode Neural Network backpropagation Algoritma Conjugate Gradient. *Jurnal Eltikom*, *1*, 2

[9] Faustina, R. S., Agoestanto, A., & Hendikawati, P. (2017). Model Hybrid ARIMA-GARCH Untuk Estimasi Volatilitas Harga Emas Menggunakan Software R. *UNNES Journal of Mathematics*, *6*(1), 11-24

[10] Zhu, Y., & Zhang, C. (2018). Gold price prediction based on pca-ga-bp neural network. *Journal of Computer and Communications*, *6*(7), 22-33.

[11] Azam, D. F., Ratnawati, D. E., & Adikara, P. P. (2018). Prediksi Harga Emas Batang Menggunakan Feed Forward Neural Network Dengan Algoritme Genetika. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, *2548*, 964X

[12] Hadavandi, E., Shavandi, H., & Ghanbari, A. (2010, August). A genetic fuzzy expert system for stock price forecasting. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery* (Vol. 1, pp. 41-44). IEEE.

[13] Rahmawati, N. E. (2012). Prediksi Data Time Series Menggunakan Fuzzy Inference System dan Particle Swarm Optimization (Studi Kasus: Prediksi Harga Emas)

[14] Rusbariand, S. P., Masodah, R., & Herawati, S. (2012). Analisis pengaruh tingkat inflasi, harga minyak dunia, harga emas dunia, dan kurs rupiah terhadap pergerakan jakarta islamic index di bursa efek indonesia. In *Prosiding Seminar Nasional* (Vol. 1, pp. 724-740).

[15] SYARAT-SYARAT, U. M. S., SATU, M. G. S. S., & ADIB, A. M. (2009). Pengaruh Inflasi, Suku Bunga Domestik, Suku Bunga Luar Negeri dan Kurs Terhadap Indeks Harga Saham (Studi pada JII dan IHSG tahun 2005-2007).

[16] Sodiq, A. (2016). Ka jian Historis Tentang Dinar dan MaMata UaUang Berstandar Emas. *IQTISHADIA*, *8*(2).

[17] Iman, N. (2009). Investasi Emas. *Jakarta: Daras Books*.

[18] Indriasari, T. (2011). Pengaruh harga minyak dunia, nilai tukar rupiah dan tingkat suku bunga SBI terhadap Jakarta Islamic Index (JII). *Pengaruh harga minyak dunia, nilai tukar rupiah dan tingkat suku bunga SBI terhadap Jakarta Islamic Index (JII)/Titik Indriasari*

[19] Sidarta, W. (2010). Pengaruh gejolak harga minyak mentah terhadap IHSG

[20] Muhammad Wildan, M. W. (2016). *PRODUK MURABA> HA> H LOGAM INVESTASI ABADI DI PEGADAIAN SYARIAH PERSPEKTIF HUKUM ISLAM (Studi Kasus di PT. Pegadaian Syariah Cabang Purwokerto)* (Doctoral dissertation, IAIN PURWOKERTO)

[21] Witjaksono, A. A. (2010). *Analisis Pengaruh Tingkat Suku Bunga SBI, Harga Minyak Dunia, Harga Emas Dunia, Kurs Rupiah, Indeks Nikkei 225, dan Indeks Dow Jones terhadap IHSG (studi kasus pada IHSG di BEI selama periode 2000-2009)* (Doctoral dissertation, UNIVERSITAS DIPONEGORO)

[22] Christina, C. (2012). Prediksi Harga Emas Menggunakan Metode Neuro-Fuzzy Tipe 2. *Telkom University, Bandung*

Genetic Algorithm Optimization on Naive Bayes for Airline Customer Satisfaction Classification

Donny Maulana[1 *)], Yoga Religia[2]
[1,2]Informatics Engineering Study Program, Faculty of Engineering, Pelita Bangsa University
email: [1]donny.maulana@pelitabangsa.ac.id , [2]yoga.religia@pelitabangsa.ac.id

Abstract –Airline companies need to provide satisfactory service quality so that people do not switch to using other airlines. The way that can be used to determine customer satisfaction is to use data mining techniques. Currently, the website www.kaggle.com has provided Airline Passenger Satisfaction data consisting of 22 attributes, 1 label and 25976 instances which are included in the supervised learning data category. Based on several previous studies, the Naïve Bayes algorithm can provide better classification performance than other classification algorithms. Several studies also state that the use of Naive Bayes can be optimized using Genetic Algorithm (GA) to obtain better performance. The use of Genetic Algorithm for Nave Bayes optimization in classifying Airline Passenger Satisfaction data requires further research to ensure the performance of the given classification. This study aims to compare the use of the Naive Bayes algorithm for the classification of Airline Passenger Satisfaction with and without GA optimization. The data validation process used in this study is to use split validation to divide the dataset into 95% training data and 5% testing data. The test results show that the use of GA on Naive Bayes can improve the classification performance of Airline Passenger Satisfaction data in terms of accuracy and recall with an accuracy value of 85.99% and a recall of 87.91%.

Keywords - data mining, classification, Naïve Bayes, Genetic Algorithm, Customer Satisfaction.

I. INTRODUCTION

Geographically, Indonesia, which is an archipelagic country, requires transportation facilities that make it easier for people to accommodate accommodation, one of which is by air. This is a great potential that can be taken by airline companies [1]. Airline companies need to provide satisfactory service quality so that people do not switch to using other airlines [2]. The service quality of an airline cannot be measured from the company's point of view, but must be seen from the point of view of customer satisfaction [3]. The method that can be used to determine customer satisfaction is to use data mining techniques [4].

One way that can be used to predict customer satisfaction with data mining techniques is by using a classification model. Classification models can be used on supervised learning data [5]. Currently on the websitewww.kaggle.com has provided Airline Passenger Satisfaction data consisting of 22 attributes, 1 label and 25976 instances included in the supervised learning data category [6], so that it can be used to create a classification model. It takes a good algorithm for making an optimal classification model, one of which uses the Naïve Bayes algorithm.

Based on several previous studies, the Naïve Bayes algorithm can provide better classification performance than other classification algorithms such as k-NN, C4.5, Decision Tree, and even Neural Networks. [7] [8] [9]. These studies try to compare the Naive Bayes algorithm with classification algorithms to predict various types of datasets to find out which algorithm has the best performance. Besides being able to provide good classification performance, the Naïve Bayes algorithm can also be used for imbalance data [10] [11], so it is suitable to be used to classify Airline Passenger Satisfaction data.

Although Nave Bayes has shown outstanding classification accuracy, currently independent assumptions are rarely discussed in the Nave Bayes classification. One way to try independent assumptions in the Naïve Bayes algorithm is by attribute weighting [12]. This is also supported by Liangxiao Jiang (2019) which states that it is necessary to propose an attribute weighting method to reduce independent assumptions [13]. Attribute weighting can be done using Genetic Algorithm (GA) through Feature Selection [14].

GA is one of the optimization algorithms created to mimic some of the processes observed in natural evolution [15]. The optimization carried out by GA is to predict the right number of iterations, so that there is no need to calculate the number of different iterations to get complete occurrences of independent paths. [16]. The most significant advantage of GA is its ability to search globally as well as adaptability to a wide spectrum of problems [17]. Based on several previous studies, it is stated that the use of GA can improve the classification performance of Naïve Bayes [18] [19].

Based on previous research, it shows that GA is able to improve classification performance on Naïve Bayes, but has not found the application of GA to Naïve Bayes for the classification of airline customer satisfaction. This study analyzes GA optimization on Naïve Bayes for the classification of Airline Passenger Satisfaction data.

II. RESEARCH METHODOLOGY

A. Data used

This study uses Airline Passenger Satisfaction data taken from the sitewww.kaggle.com on April 24, 2021 [6]. Airline Passenger Satisfaction data is data that contains a survey of airline passenger satisfaction in the world. Airline Passenger Satisfaction data is still a new dataset that has not been widely used for research because the data has been uploaded to the site www.kaggle.com since May 2020. This data has 1 label with a boolean data type consisting of 22 attributes and 25976 instances. The purpose of using this data is to find out what factors are most correlated with airline passenger satisfaction, so that this data is suitable to be used to create a classification model. Each attribute and label contained in the Airline Passenger Satisfaction data can be seen in Table 1.

Table1. Airline Passenger Satisfaction Attributes and Labels

Content	Information	Ket
Gender	Passenger gender (Female, Male)	Attribute
Customer Type	Type of customer (Loyal customers, disloyal customers)	Attribute
age	Actual passenger age	Attribute
Type of Travel	Passenger flight destinations (Private Travel, Business Trip)	Attribute
Class	Class of travel on passenger aircraft (Business, Eco, Eco Plus)	Attribute
flight distance	Flight distance of this trip	Attribute
Inflight wifi service	Satisfaction level of inflight wifi service (1-5)	Attribute
Arrival time convenient	Satisfaction level Departure / Arrival time comfortable (1-5)	Attribute
Ease of Online booking	Online order satisfaction level (1-5)	Attribute
Gate location	Gate location satisfaction level (1-5)	Attribute
Food and drink	Food and beverage satisfaction level (1-5)	Attribute
Online boarding	Online boarding satisfaction level (1-5)	Attribute
Seat comfort	Seat comfort level of satisfaction (1-5)	Attribute
Inflight entertainment	Satisfaction level of inflight entertainment (1-5)	Attribute
On-board service	On-board service satisfaction level (1-5)	Attribute
Leg room service	Room service satisfaction level (1-5)	Attribute
Baggage handling	Baggage handling satisfaction level (1-5)	Attribute
Check-in service	Check-in service satisfaction level (1-5)	Attribute
Inflight service	In-flight service satisfaction level (1-5)	Attribute
Cleanliness	Cleanliness satisfaction level Tingkat (1-5)	Attribute
Departure Delay	Minutes delayed on departure	Attribute
Arrival Delay	Minutes delayed on Arrival	Attribute
Satisfaction	Airline satisfaction level (Satisfied, Dissatisfied)	Label

Airline Passenger Satisfaction Data does not have a missing value, so it can be directly used for the classification process without the need to go through preprocessing data.

B. Research Model

Airline Passenger Satisfaction data is used to form a classification model. The label used is the attribute "Satisfaction" with a value of "Satisfied" and "Unsatisfied". From all data used, 66% are instances labeled "Not Satisfied" while the rest are instances labeled "Satisfied". This research carried out the test twice which later will be analyzed the results obtained. The first test is done using GA optimization, while the second test is done without GA optimization.

The classification model built in this study uses the spit validation process to divide the data into training data and testing data. The training data used in this study is 95% of all Airline Passenger Satisfaction data, while the remaining 5% is used for testing data. The training data obtained from the validation process will be used for classification modeling using the Naïve Bayes algorithm. The resulting model is then used as an apply model for use in testing data. After the classification has been carried out, then the performance of the classification model is measured based on the values of accuracy, precision, and recall.

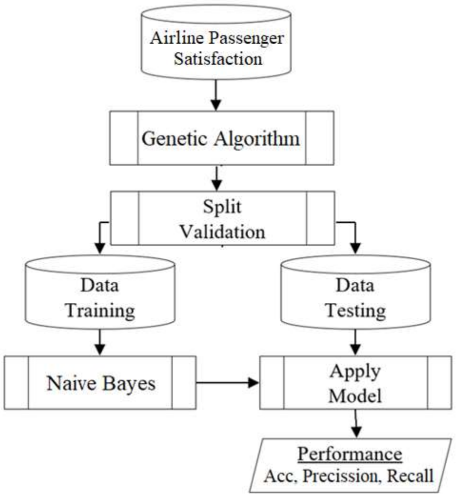


Figure 1. First Test of Naïve Bayes Classification Using Genetic Algorithm

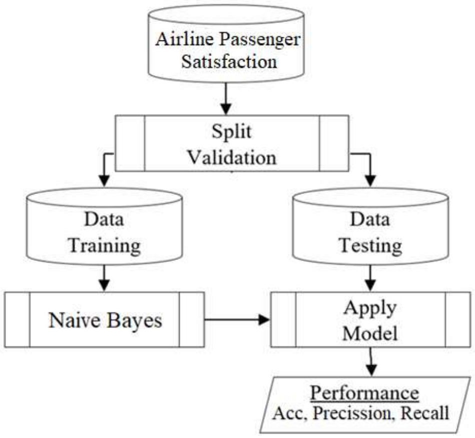


Figure 2. First Test of Naïve Bayes Classification Without Using Genetic Algorithm

In Figure 1 and Figure 2 shows that in this study the test was carried out 2 times, namely: (1) Classification of Airline Passenger Satisfaction data using Naïve Bayes with optimization of Genetic Algorithm, (2) Classification of Airline Passenger Satisfaction data using Naïve Bayes without optimization of Genetic Algorithm . The performance results of the two tests will be compared and then analyzed to show the research findings.

C. Classification with Naïve Bayes

Naïve Bayes is widely used to solve classification problems in real-world applications because of its ease of building and interpreting data, and its good performance. [13]. The Naïve Bayes algorithm is a supervised learning algorithm based on the Bayes theorem with the assumption of independence between predictors. This means that the features in the class are independent of other features. The Naive Bayes classifier can be used for both continuous and categorical variables [12]. It is based on the Bayes formula which is the probability of event A given proof of B which can be seen in the following equation [7]:

$$P(A,B) = P(A)P(B) \qquad (1)$$

Through equation (1) and using the concept of the Bayes theorem, the final equation of the Naïve Bayes algorithm is obtained as follows:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad (2)$$

Based on equation (2), it is known that A is a class and B is an instance. A represents the dependent event which means the predicted variable and B represents the previous event which means the predictor attribute. The final step of the Naive Bayes algorithm is to find the maximum probability that will serve as a predictor class.

D. Optiomation with Genetic Algorithm

Genetic Algorithm (GA) was discovered by John Holland in 1960 who was inspired by the process of evolution in nature [20]. GA is an optimization method developed based on the mechanism of natural selection by imitating the genetics of living things in solving difficult problems with high complexity and undesirable structures. [21]. The optimization process in GA is carried out based on the sample population by developing a population candidate solution towards a better solution [22].

The first step of GA is the formation of chromosomes. Each chromosome yields one answer to one problem. New answers are generated after applying the crossover, mutation, and selection operations. The fitness function evaluates the benefits of chromosomes. GA then finds the most feasible chromosome with the maximum fitness function value from generation to generation. Many circumstances such as initial population size, number of generations, crossover operator, mutation operator and fitness function determine the performance of the genetic algorithm [23]. Fewer generations are required to reach the optimal answer in order to produce a more accurate fitness function.

E. Evaluation with Cross Validation

The cross validation method or also known as k-fold cross validation is a validation method that involves splitting a random sample set into a series of equal-sized folds (groups), where k indicates the number of partitions, or folds, the data set is broken down. [24]. For example, if the k value of ten is used, the data set is divided into ten partitions. In this case, nine partitions are used for training data, while the other partitions are used for data testing. The training is repeated ten times, each time using a different partition as the test set, then the other nine partitions are used as training data. The results are then averaged for reporting [25].

F. Confution Matrix for Performance Testing

In a binary confution matrix, observations that are correctly classified into a positive class are called true positives (TP) and observations that are correctly classified into a negative class are called true negatives (TN). Instances of a positive class that are classified incorrectly as negative are called false negatives (FN) and instances of a negative class that are classified incorrectly as positive are called false positives (FP). Based on the values of TP, FP, TN and TP, classification performance indicators can be calculated that reflect how the classifier performs in detecting a given class. The most commonly used indicators are accuracy, precision, recall (sensitivity) which can be written in the following equation [26]:

$$Akurasi = \frac{TP+TN}{(TP+TN+FP+F\ \ )} \, (3)$$

$$Presisi = \frac{TP}{(TP+FP)} \, (4)$$

$$Recall = \frac{TP}{(TP+F\ \ )} \, (5)$$

Accuracy is the simplest and most widely used metric for measuring the performance of a classification model. In addition to using accuracy, this study also considers classification performance measures in terms of precision and recall. According to Brendan Juba and Hai S. Le (2019), classification performance measures using accuracy, precision and recall are recommended because they are suitable for classification of imbalance data. [27].

III. RESULTS AND DISCUSSION

A. Testing Step

The Rapid Miner version 5.0 tools were used in this study to conduct testing. Rapid Miner can be used for research, rapid prototyping, and supports all steps of the data mining process such as data preparation, result visualization, validation and optimization. [28], so it is considered suitable for use in this study. The first stage in making a research model is to call the data *Airline Passenger Satisfaction*Rapid Miner tools, then the multiply function is performed to perform two tests at once, namely testing using GA and testing without using GA. The data validation process is carried out using split validation to divide the data into 95% training data and 5% testing data. In more detail about the data calling and validation process can be seen in Figure 3.

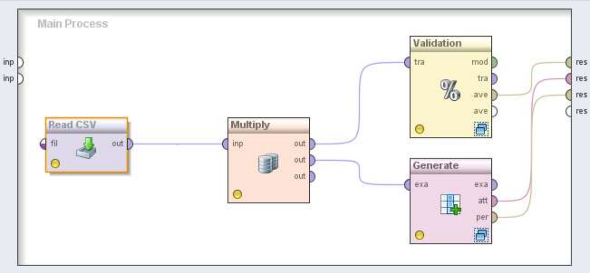


Figure 3. Data Calling and Validation Process

In each validation process shown in Figure 3, it contains a learning process with the Naïve Bayes algorithm which is then applied to the apply model to measure the performance of accuracy, precision and recall. The learning process in this study can be seen in Figure 4.

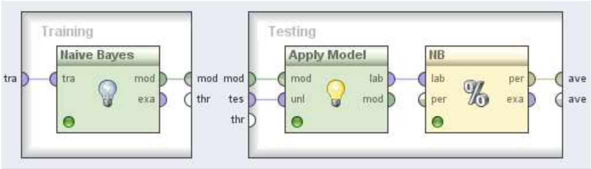


Figure 4. The Learning Process of Naïve Bayes and Apply Model

The next step after all research models have been formed is to run the model that has been built on Rapid Miner, then the results of accuracy, precision and recall will be obtained for analysis of the results.

B. Test result

After 2 tests, the accuracy, precision, and recall values of the two models were obtained. More complete test results can be seen in Table 2.

Table2. Test result

No.	Algorithm	Accuracy	Precision	Recall
1	Naive Bayes	84.53%	88.47%	84.90%
2	GA + Naive Bayes	85.99%	87.43%	87.91%

Based on Table 2, it can be seen that GA is able to improve the accuracy and recall of Naïve Bayes, but GA has not been able to increase the precision value of Naïve Bayes. The test results show that with an accuracy of 85.99%, GA optimization gives Naïve Bayes an increase in accuracy value of 1.46% and an increase in recall value of 3.01% for Airline Passenger Satisfaction data classification.

Table3. Genetic Algorithm Weighting Results on Data*Airline Passenger Satisfaction*

Attribute	Weighting
Gender	0
Customer Type	0
age	0
Type of Travel	0
Class	1
Flight Distance	0
Inflight wifi service	1
Departure / Arrival time convenient	0
Ease of Online booking	0
Gate location	0
Food and drink	0
Online boarding	0
Seat comfort	0
Inflight entertainment	0
On-board service	1
Leg room service	0
Baggage handling	0
Checkin service	1
Inflight service	0
Cleanliness	0
Departure Delay in Minutes	0
Arrival Delay in Minutes	0

However, it turns out that the use of GA also reduces the precision value by 1.04% from the use of Naïve Bayes for the classification of Airline Passenger Satisfaction data. This is presumably because of the 22 attributes in the Airline Passenger Satisfaction data, it turns out that only 3 attributes are weighted by GA. This weighting result also explains why the increase in accuracy and recall provided by GA is not too large. In Table 3, it can be seen that there are only 4 attributes that are weighted by GA. This shows that, based on the 4th GA, these attributes are the most important to consider when classifying Airline Passenger Satisfaction data. The attributes are: Class, Inflight wifi service, On-board service and Check-in service.

C. Discussion of Results

Based on the results of the tests that have been carried out, classification Airline Passenger Satisfaction data has shown that the use of GA optimization can improve the accuracy and recall performance of the Naïve Bayes algorithm, although not too large. The small increase in performance given is thought to be because the attributes given weighting by GA are less than 25% of all the attributes in the Airline Passenger Satisfaction data. This makes the probability calculation process in Naïve Bayes less influential. Even in terms of precision, it turns out that the use of GA actually decreases the performance of Naïve Bayes.

Although the optimization of GA does not give maximum results, by using GA it turns out which attributes can be obtained which can be used as evaluation priorities to see the satisfaction of airline customers. By looking at the attributes given weighting by GA, it can be used as a reference to consider these attributes as the main focus for service improvement. The attributes that are given weighting by GA include: Class, Inflight wifi service, On-board service and Checkin service. This finding is expected to provide a practical contribution to the future services that will be provided by airlines to their customers.

VI. CONCLUSION

This study has tested the use of the Naïve Bayes algorithm to classify Airline Passenger Satisfaction data and compared it with the Nave Bayes classification using GA optimization. Based on the tests that have been carried out, it shows several results, namely:

1. The highest accuracy and recall of Airline Passenger Satisfaction data classification is using the Naïve Bayes algorithm with GA optimization. The maximum accuracy obtained is 85.99% and the maximum recall is 87.91%.
2. The maximum precision value from the classification of Airline Passenger Satisfaction data is to use the Naïve Bayes algorithm without GA optimization with a precision value of 88.47%.
3. The GA algorithm has not been able to provide maximum performance addition to the Naïve Bayes algorithm to classify Airline Passenger Satisfaction data.
4. Attributes Class, Inflight wifi service, On-board service and Checkin service are attributes that need to be considered by airlines to maximize customer satisfaction.

The results of this study are still not able to provide a good enough performance for Airline Passenger Satisfaction data classification, because neither accuracy, precision nor recall has a score of more than 90%. This requires further research to obtain a better Airline

Passenger Satisfaction data classification model in the future. Based on the findings of this study, it is suggested that future research can apply other optimization methods to further optimize the performance of the Naïve Bayes algorithm, for example the Particle swarm optimization (PSO) algorithm or boostraping.

REFERENCES

[1] E. L. Widjaja, A. Aprilia dan A. Harianto, "Analisa Pengaruh Kualitas Layanan Terhadap Kepuasan Penumpang Maskapai Penerbangan Batik Air," *Jurnal Hospitality dan Manajemen Jasa,* vol. 5, no. 2, pp. 118-132, 2017.

[2] W. Ardhia, "Tingkat Kepuasan Penumpang Terhadap Layanan Maskapai Penerbangan PT. Lion Air Rute Menuju Jakarta," *Jurnal Perhubungan Udara,* vol. 41, no. 1, pp. 19-28, 2015.

[3] M. D. Darus dan K. Mahalli, "Analisis Tingkat Kepuasan Penumpang Terhadap Kualitas Pelayanan di Bandar Udara Internasional Kualanamu," *Jurnal Ekonomi dan Keuangan,* vol. 3, no. 6, pp. 408-420, 2015.

[4] M. S. Garver, "Using Data Mining for Customer Satisfaction Research," *Marketing Research,* vol. 14, no. 1, pp. 8-17, 2002.

[5] V. Gopalakrishnan dan C. Ramaswamy, "Patient Opinion mining to Analyze Drugs Satisfaction Using Supervised Learning," *Journal of Applied Research and Technology,* vol. 15, no. 1, pp. 311-319, 2017.

[6] Kaggle, "Kaggle.com," Mei 2020. [Online]. Available: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction. [Diakses 24 Maret 2021].

[7] I. A. A. Amra dan A. Y. A. Maghari, "Students Performance Prediction Using KNN and Naïve Bayesian," dalam *8th International Conference on Information Technology (ICIT)*, Al-Zaytoonah University of Jordan, Jordan, 2017.

[8] F. Osisanwo, J. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi dan J. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison," *International Journal of Computer Trends and Technology (IJCTT),* vol. 48, no. 3, pp. 128-138, 2017.

[9] E. N. Azizah, U. Pujianto, E. Nugraha dan Darusalam, "Comparative Performance Between C4.5 and Naive Bayes Classifiers in Predicting Student Academic Performance in A Virtual Learning Environment," dalam *4th International Conference on Education and Technology (ICET)*, Malang, Indonesia, 2018.

[10] K. Madasamy dan M. Ramaswami, "Data Imbalance and Classifiers: Impact and Solutions from A Big Data Perspective," *International Journal of Computational Intelligence Research,* vol. 13, no. 9, pp. 2267-2281, 2017.

[11] E. M. Hassib, A. I. El-Desouky, E.-S. M. El-Kenawy dan S. M. El-Ghamrawy, "An Imbalanced Big Data Mining Framework for Improving Optimization Algorithms Performance," *Journal & Magazines,* vol. 7, no. 1, pp. 170774-170795, 2019.

[12] S. Chen, G. I. Webb, L. Liu dan X. Ma, "A Novel Selective Naïve Bayes Algorithm," *Knowledge-Based Systems,* vol. 192, pp. 1-15, 2020.

[13] L. Jiang, L. Zhang, L. Yu dan D. Wang, "Class-Specific Attribute Weighted Naive Bayes," *Pattern Recognition,* vol. 88, no. 1, pp. 321-330, 2019.

[14] S. Ernawati, R. Wati, N. Nuris, L. S. Marita dan E. R. Yulia, "Comparison of Naïve Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application," *Journal of Physics: Conference Series,* vol. 1641, pp. 1-7, 2020.

[15] S. Ernawati, E. R. Yulia, Frieyadie dan Samudi, "Implementation of The Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies," dalam *The 6th International Conference on Cyber and IT Service Management (CITSM 2018)*, Medan, Indonesia, 2018.

[16] A. Arwan dan D. S. Rusdianto, "Optimization of Genetic Algorithm Performance Using Naïve Bayes for Basis Path Generation," *Kinetik,* vol. 2, no. 4, pp. 273-282, 2017.

[17] E. Stripling, S. v. Broucke, K. Antonio, B. Baesens dan M. Snoecka, "Profit Maximizing Logistic Model for Customer Churn Prediction Using Genetic Algorithms," *Swarm and Evolutionary Computation,* vol. 40, no. 1, pp. 116-130, 2018.

[18] D. K. Choubey, S. Paul, S. Kumar dan S. Kumar, "Classification of Pima Indian Diabetes Dataset Using Naive Bayes With Genetic Algorithm As An Attribute Selection," dalam *The International Conference on Communication and Computing Systems (ICCCS)*, Ranchi, India, 2016.

[19] L. G. P. Suardani, I. M. A. Bhaskara dan M. Sudarma, "Optimization of Feature Selection Using Genetic Algorithm with Naïve Bayes Classification for Home Improvement Recipients," *International Journal of Engineering and Emerging Technology,* vol. 3, no. 1, pp. 66-70, 2018.

[20] M. Melanie, An Introduction to Genetic Algorithms, London, England: First MIT Press, 1999.

[21] Y. Religia, A. Nugroho dan W. Hadikristanto, "Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," *Jurnal Rekayasa Sistem dan Teknologi Informasi,* vol. 5, no. 1, pp. 187-192, 2021.

[22] E. Habibi, M. Salehi, G. Yadegarfar dan A. Taheri, "Optimization of ANFIS Using A Genetic Algorithm for Physical Work Rate Classification," *International Journal of Occupational Safety and Ergonomics,* vol. 26, no. 3, pp. 436-443, 2020.

[23] H. Motieghader, A. Najafi, B. Sadeghi dan A. Masoudi-Neja, "A Hybrid Gene Selection Algorithm for Microarray Cancer Classification Using Genetic Algorithm and Learning Automat," *Informatics in*

Medicine Unlocked, vol. 9, no. 1, pp. 246-254, 2017.

[24] C. A. Ramezan, T. A. Warner dan A. E. Maxwell, "Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification," *Remote Sensing,* vol. 11, no. 2, pp. 2-21, 2019.

[25] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society,* vol. 36, no. 2, pp. 111-147, 1974.

[26] S. Ruuskaa, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen dan J. Mononen, "Evaluation of The Confusion Matrix Method in The Validation of An Automated System for Measuring Feeding Behaviour of Cattle," *Behavioural Processes,* vol. 148, no. 1, pp. 56-62, 2018.

e-ISSN: 2614-8404
p-ISSN : 2776-3234

Assessment of Employee Using Simple Multi-Attribute Technique Exploiting Rank (SMARTER) and Behaviorally Anchor Rating Scale (BARS) Method

Heni Sulastri[1 *)]
[1] Informatics Department of Engineering Faculty, Siliwangi University
Tasikmalaya, Indonesia
email: [1] henisulastri@unsil.ac.id

Abstract –Lecturers' active role as the spearhead of higher education has an essential role in improving higher education quality and sustainability. Therefore, assessing work behaviour is needed to measure how lecturers participate in achieving the vision and mission, quality improvement, and service guarantee to students and complementary documentation. This condition became the basis of research. They are implementing decision support systems with Simple Multi-Attribute Rating Technique Exploiting Ranges (SMARTER) and Graphic Rating Scale (GRS) to measure a lecturer's behaviour by using multiple criteria. With the SMARTER method and Behaviorally Anchor Rating Scale (BARS). By applying the impermeable BARS method, the work behaviour assessment process results in ease and accuracy that is more in line with the employees' behaviour being assessed. With the SMARTER approach, an assessment of employee work behaviour is produced, with 90% of alternatives used. The results are Good.

Keywords - Lecturer, BARS method, Method of SMARTER, Behavioral Assessment Work..

I. INTRODUCTION

Human resources have an essential role in the sustainability of an agency. Higher education is one of the educational institutions that have lecturers as human resources where lecturers' presence is one of the factors that is considered absolute. Lecturers are prominent supporters who interact directly with students. A lecturer is deemed qualified if he meets the qualifications and work behaviour and is competent in line with its vision and mission. Success is usually measured by the lecturer's level of success in teaching, the level of discipline in education, the ability to interact with students, and many other supporting factors [17].

In any organization, Behavior Assessment or the performance of each employee is an everyday activity. As stated by [7] which states that employee Behavior Assessment can be said to be effective if it includes the following two things, namely (1) the existence of a set of standards and (2) information communication (feedback). Dessler [10] "Effective appraisal also requires that the supervisor set performance standards. And it requires that the employee receives the training, feedback, and incentives required to eliminate performance deficiencies". Gary Dessler's opinion is increasingly confirmed that Conduct's assessment effectively requires a standard that has been I preset and feedback to prevent a decline. Likewise, in higher education institutions, whether in the form of universities, institutes, or colleges. In general, Job Performance Appraisal is a process by which organizations evaluate performance to improve performance [7].

Assessment of lecturer achievement aims to achieve the vision and mission of higher education institutions and accreditation needs [12].

Decision-making methods are used to be applied for job performance assessment. One of them is the Simple Multi-Attribute Rating Technique Exploiting Ranges (SMARTER) method, which supports multi-criteria by giving weight to each criterion and sub-criteria that illustrates how critical the requirements are [2][10][13][18]. Each standard and sub-criteria, which are characteristics or several properties of items or items, will be presented by applying the *Behaviorally Anchor Rating Scale* (BARS) method [11][15].

This article proposes the combination of Simple Multi-Attribute Rating Technique Exploiting Ranges (SMARTER) method and Behaviorally Anchor Rating Scale (BARS) to analyze the lecturer's performance in Siliwangi University.

A. Job Performance Assessment

Job Performance Appraisal is a formal system for assessing and evaluating the performance of an individual or team assignments used by industry, agencies, and organizations to generate feedback on performance following the standard set used [1][5]. Correct Job Performance Assessment will help relevant stakeholders and the employees or Human Resource Development division being assessed. The Job Performance Appraisal process consists of three stages: (1) defining the job, evaluating performance, and providing feedback [1].

B. Decison Support System

Decision support systems were first put forward in the early 1970s by Michael S. Scott Morton. It was term Management Decision Systems to assist managers in making decisions on semi-structured problems, providing support for managers, increasing managers' decisions, speed computing, and productivity enhancement [10][14]. Decision support systems are considered capable of solving problems and solving semi-structured issues [13]. A semi-structured problem is a problem that includes several elements recognized by problem solvers. Decision-making correlates with the uncertainty of the results of the decisions taken to reduce risk factors.

The decision-making process consists of three phases, namely (1) the Intelligent step or the operation of tracking and detecting problems and identifying problems; (2) the design phase or the phase to understand the problem, reduce the risk and test the feasibility of the risk by conducting a process of finding, developing and analyzing alternative actions that can be taken; (3) Choice or a decision-making process based on the implemented alternative [3].

C. Simple Multi-Attribute Rating Technique Exploiting Ranges (SMARTER)

SMART is a multi-criteria decision-making method. The multi-criteria decision-making technique is based on the theory that each alternative consists of several criteria that have value - value. Each standard has a weight that describer how important criteria are compared with other criteria [4]. The SMARTER method is developing the Simple Multi-Attribute Rating Technique (SMART) method introduced by Edward in 1977 [10]. In the SMARTER method, the Rank Order Centroid (ROC) weighting formula is used [6].

The equation for the SMARTER method can be seen in the following equation (1), where $U_n = Final\ Score$, $W_k = Weighting\ from\ criteria\ k$, $U_n(X_{nk})$ =The utility value for the k criterion for the k alternative.

$$U_n \ = \ \sum_{k=1}^{k} W_k U_n \left(X_{nk}\right) \tag{1}$$

Calculation of utility value can use the following equation (2), where $U_i(a_i)$ is utility value for *(i)* criteria, C_i is the value of the *(i)* criteria, C_{min} is the minimum value of criteria, C_{max} is the maximum value of criteria.

$$U_i(a_i) = \ 100\% \ \times \frac{(C_i - C_{min})}{(C_{maks-} \, C_{min})} \tag{2}$$

D. Weighting Rank Order Centroid (ROC)

The ROC technique's weighting works by giving weight to the criteria according to the ranking based on the priority level. The weighting of the ROC is generally formulated in equation (3), where W is the weight value of criteria, k is the number of criteria and i is alternative value.

$$W_k = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{i}\right) \tag{3}$$

E. Behavioural Anchor Rating Scale (BARS)

The Behavioral Anchor Rating Scale (BARS) method is a performance appraisal method that combines work behaviour approaches with personal traits. Scaling is done between 5 to 10 vertical actions (Anchor) for each work indicator. Anchors are arranged from the highest value to the lowest cost. Anchors can be in the form of critical incidents obtained through job analysis, usually compiled by a team of Human Resources specialists, managers, and employees [8]. The stages in the *Behavioral Anchor Rating Scale (BARS)* method can be seen as follows:

a) Making a Critical Incident
b) Developing performance dimensions
c) Reallocating events
d) Making the scale of the incident
e) Developing the final tools

The BARS method has several positive values that are more accurate because the experts have developed the BARS in the HRD devising. HRD has more precise standards, can generate feedback, systematically critical group events (Anchors) into five to ten independent dimensions, and has consistent properties [8][11] [15].

F. Related Research

Several studies on the SMARTER and BARS methods have been carried out to optimize of decision-maker. Alfa Saleh et al. in 2018 determine the selection of laboratory assistants by applying the Simple Multi-Attribute Rating Technique Exploiting Ranges (SMARTER) method by applying six criteria and weighting accordingly [10]. With assessment priorities and produce research results that the technique used can provide useful recommendations. Other related research involves the Simple Multi-Attribute Rating Technique Exploiting Ranges (SMARTER) method to determine life insurance product recommendations to customers. Research results show that the SMARTER method is optimal and feasible as alternative decision support by Haryanti et.all in 2016 [6]. This research is also strengthened by other research that applies the SMARTER method in selecting and evaluating suppliers of Brazil's construction industry. The SMARTER method is considered efficient in selecting suppliers, providing supplier recommendations in the form of ranking by prioritizing the quality and price offered by each supplier by Schram and Danielle in 2012 [9]

Related research regarding the BARS method includes a study conducted by Michelle Martin-Raugh, et al. 2016 [8] regarding the application to evaluate teaching practice with the results of her research stating that the BARS method is preferred the assessment process than the FFT method. Other research related to BARS, such as that conducted by Donald P Schwab et al. in 2006 [11], measured BARS with the following three characteristics: Leniency Effect, Independent Dimension, and Reliability results showing that the BARS method still needs further research.

II. RESEARCH METHODOLOGY

The stages in the research are carried out as in the following figure:

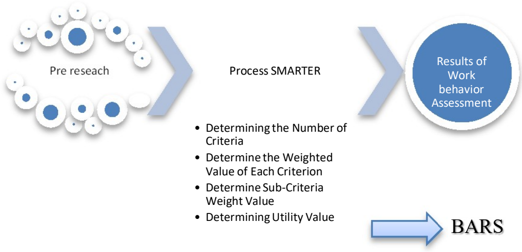

Pre reseach
Process SMARTER
Results of Work behavior Assessment
Determining the Number of Criteria
Determine the Weighted Value of Each Criterion
Determine Sub-Criteria Weight Value
Determining Utility Value
BARS

Figure 1. Research Diagram

This research is divided into three main stages, namely (1) Pre research, (2) process SMARTER, (3) behaviour Assessment System.

A. Pre Research

The pre-research begins with direct and indirect observations of the lecturer job performance assessment in one of the tertiary institutions to find an overview of the Lecturer Job Performance Assessment process. The resulting observations' results are continued by identifying the problem to produce a problem formulation and limitation. A literature study is conducted to provide research guidance in finding solutions to solve the problems formulated.

B. Process SMARTER

SMARTER method that is carried out consists of the following stages:

a) Determining the Number of Criteria

The Lecturer Work Behavior Assessment Process that is carried out refers to Law No. 5 of 2014, Articles 75-78 of ASN, and Government Regulation No.46 of 2011 concerning Assessment of Civil Servant Work Performance. The criteria used are as follows:

Table 1. Criteria

No.	Criteria	The Type of Criteria
1	Service Orientation	Categorical
2	Integrity	Categorical
3	Commitment	Categorical
4	Discipline	Categorical
5	Cooperation	Categorical
6	Leadership	Categorical

b) Determine the Weight Value of Each Criterion

Each criterion's weight and priority levels are determined based on the priority level using equation 3, namely Rank Order Centroid (ROC) weighting.

Table 2 Weighted Criteria Value

No.	Criteria	Priority Level	Weighted Value
1	Service Orientation	1	0.408
2	Integrity	2	0.242
3	Commitment	3	0.158
4	Discipline	4	0.103
5	Cooperation	5	0.061
6	Leadership	6	0.028

Based on table 2 above, Service Orientation has the highest weight value according to the priority level enforced on the grounds that lecturers have priority to provide services both internally and externally.

c) Weight Value of Sub Criteria

Table 3. Sub-criteria weight values

No.	Criteria	Sub-criteria	Weight
1	Service Orientation	$91 \leq$ score ≤ 100	0.457
		$76 \leq$ score ≤ 90	0.257
		$61 \leq$ score ≤ 75	0.157
		$51 \leq$ score ≤ 60	0.090
		Under 50	0.040
2	Integrity	$91 \leq$ score ≤ 100	0.457
		$76 \leq$ score ≤ 90	0.257
		$61 \leq$ score ≤ 75	0.157
		$51 \leq$ score ≤ 60	0.090
		Under 50	0.040
3	Commitment	$91 \leq$ score ≤ 100	0.457
		$76 \leq$ score ≤ 90	0.257
		$61 \leq$ score ≤ 75	0.157
		$51 \leq$ score ≤ 60	0.090
		Under 50	0.040
4	Discipline	$91 \leq$ score ≤ 100	0.457
		$76 \leq$ score ≤ 90	0.257
		$61 \leq$ score ≤ 75	0.157
		$51 \leq$ score ≤ 60	0.090
		Under 50	0.040
5	Cooperation	$91 \leq$ score ≤ 100	0.457
		$76 \leq$ score ≤ 90	0.257
		$61 \leq$ score ≤ 75	0.157
		$51 \leq$ score ≤ 60	0.090
		Under 50	0.040
6	Leadership	$91 \leq$ score ≤ 100	0.457
		$76 \leq$ score ≤ 90	0.257
		$61 \leq$ score ≤ 75	0.157
		$51 \leq$ score ≤ 60	0.090
		Under 50	0.040

Based on table 3 above, the weight values for each sub-criterion are categorized based on the achievement number $91 \leq$ core ≤ 100 for the title Very Good, $76 \leq$ score ≤ 90 for the title Good, $61 \leq$ scores ≤ 75 for the term Enough, $51 \leq$ scores ≤ 60 for Less, and Under 50 for Bad designations.

While in Table 4 shows the formulation of assessment using the bars method, where the evaluation has sub-criteria.

Table 4. Formulation of Assessment using the BARS Method

Indicator	Rating	Anchor
Service Orientation	Very good	Always be able to complete service tasks and possible with a polite and very satisfying attitude for both internal and external service to the organization.
	Good	In general, can complete service tasks well with a polite and satisfying attitude for both internal and external service to the organization
	Enough	Sometimes he can complete service tasks quite well, and the attitude is quite polite and satisfying enough for both internal and external services to the organization.
	less	Not complete service tasks correctly and attitude less polite and unsatisfactory for both internal and external services.
	Bad	Not completing the task with good service and rude attitude and unsatisfactory both for the internal and external service organization.
Integrity	Very good	Always in carrying out duties, to be honest, sincere, and never abuse one's authority and dare to bear the risk of the actions he/she does.
	Good	In general, carrying out the tasks honestly, sincere, and never to abuse the authority and responsibility to the actions taken.
	Enough	Occasionally / sometimes, in carrying out his duties, he is quite honest, entirely sincere. Sometimes he misses the authority

e-ISSN: 2614-8404
p-ISSN : 2776-3234

		and is brave enough to bear the risk of his actions.
	less	Lack of honesty, lack of sincerity, carrying out their duties, and often misuse their authority, but they are not brave enough to bear the risk of their actions.
	Bad	No, dishonest, sincere, in performing the task, and always abusing his authority and did not dare to risk their actions.
Commitments	Very good	Always work diligently to uphold the ideology of the state Pancasila, 1945 Constitution Of The Republic Of Indonesia, Unitary State of the Republic of Indonesia, this singular diversity and government plans with the aim to be able to carry out its duties and prioritize the interests of the government rather than personal interests and / or groups in accordance with duties, functions and his responsibilities as a state apparatus to workplace organizations
	Good	Generally tried earnestly to uphold the ideology of the state Pancasila, 1945 Constitution Of The Republic Of Indonesia, Unitary State of the Republic of Indonesia, this singular diversity and government plans with the aim to be able to carry out its duties and prioritize the interests of the government rather than personal interests and / or groups in accordance with duties, functions and his responsibilities as a state apparatus to workplace organizations
	Enough	Sometime trying earnestly to uphold the ideology of the state Pancasila, 1945 Constitution Of The Republic Of Indonesia, Unitary State of the Republic of Indonesia, this singular diversity and government plans with the aim to be able to carry out its duties and prioritize the interests of the government rather than personal interests and / or groups in accordance with duties, functions and his responsibilities as a state apparatus to workplace organizations
	less	Less trying in earnestly to uphold the ideology of the state Pancasila, 1945 Constitution Of The Republic Of Indonesia, Unitary State of the Republic of Indonesia, this singular diversity and government plans with the aim to be able to carry out its duties and prioritize the interests of the government rather than personal interests and / or groups in accordance with duties, functions and his responsibilities as a state apparatus to workplace organizations
	Bad	Never tried earnestly to uphold the ideology of the state Pancasila, 1945 Constitution Of The Republic Of Indonesia, Unitary State of the Republic of Indonesia, this singular diversity and government plans with the aim to be able to carry out its duties and prioritize the interests of the government rather than personal interests and / or groups in accordance with duties, functions and his responsibilities as a state apparatus to workplace organizations
Discipline	Very good	Always comply with laws and regulations and / or official service regulations with a sense of responsibility and still comply with the provisions of working hours and be able to properly store and / or maintain state property entrusted to them
	Good	In general, he obeys the prevailing laws and / or official regulations with a sense of responsibility, adheres to the provisions of working hours and is able to properly store and / or maintain state property entrusted to him.
	Enough	Sometimes he obeys the prevailing laws and / or official regulations with a sense of responsibility, regards the provisions of working hours and is sufficiently capable of storing and / or maintaining state-owned goods entrusted to him quite well, and not entering or being late for work. and go home sooner than the stipulated working hours without valid reasons for 5 (five) to 15 (fifteen) working days.
	less	Lack of obeying the prevailing statutory regulations and/or official service regulations with a sense of lack of responsibility, obeying working hours regulations and being unable to store and / or maintain state property entrusted to them poorly, and not entering or being late for work and go home sooner than the stipulated working hours without a valid reason for 16 (sixteen) to 30 (thirty) working days.
	Bad	Never obeyed the rules of the Law and / or the rules of business that occur with a sense of irresponsibility, comply with the working hours and not be able to store and / or maintain state-owned goods entrusted to him in a good way, and do not enter or be late for work and return from work hours without a valid reason for more than 31 working days.
Cooperation	Very good	Always able to cooperate with colleagues, superiors, subordinates both inside and outside the organization and respect and accept the opinions of others, willing to accept decisions taken legally which have become joint decisions.
	Good	In general, they are able to cooperate with colleagues, superiors, subordinates both inside and outside the organization and respect and accept other people's opinions, are willing to accept decisions made legally which have become joint decisions.
	Enough	Sometimes able to work together with colleagues, superiors, subordinates both inside and outside the organization and sometimes respect and accept the opinions of others, sometimes willing to accept decisions taken legally which have become joint decisions.
	less	Less able to cooperate with colleagues, superiors, subordinates both inside and outside the organization and less respect and acceptance of other people's opinions, less willing to accept decisions made legally which have become joint decisions.
	Bad	Have never been able to cooperate with colleagues, superiors, subordinates both inside and outside the organization and do not respect and accept other people's opinions, are not willing to take decisions made legally which have become joint decisions.
Leadership	Very good	Always act firmly and impartially, provide a good example, the ability to move work teams to achieve high performance, capable of uplifting and moving subordinates in carrying out the task and able to make decisions quickly and accurately.
	Good	In general, act decisively and impartially, provide good role models, the ability to mobilize the work team to achieve high performance, be able to inspire and move

		subordinates in carrying out their duties and be able to make decisions quickly and accurately.
	Enough	Sometimes acting decisively and impartially, setting an example, being sufficiently capable of mobilizing the work team to achieve high performance, and enough capable of arousing enthusiasm and mobilizing subordinates in carrying out their duties and capable of making decisions quickly and accurately
	less	Lack of acting decisively and sometimes taking sides, less able to provide good role models, less able to mobilize the work team to achieve high performance, and less able to inspire enthusiasm and mobilize subordinates in carrying out tasks and less able to make decisions quickly and accurately
	Bad	Not been able to act firmly and impartially, not a good example, not be able to mobilize work teams to achieve high performance, unable to inspire the spirit and stir subordinates in carrying out their duties and are not able to make decisions quickly and accurately.

d) Determining the Value of Utilities

By applying equation 2 (two), we can get the value of Utility.

e) Work Behavior Assessment System

The final result of this lecturer work behaviour assessment can produce a ranking based on the alternatives that have been used. Ranking results can be input for superiors in making decisions.

III. RESULT AND DISCUSSION

In this study, ten lecturers' data were used to carry out work performance assessments including data on the value of Service Orientation (C1), Integrity (C2), Commitment (C3), Discipline (C4), Cooperation (C5) and Leadership (C6).

The data used from lecturers who have carried out work performance assessments for 2019 even semester, where ten lecturer data will be used as an alternative in testing the SMARTER and BARS methods. The following ten lecturer data can be seen in table 5:

Table 5. Sub-criteria weights for each alternative.

A	C1	C2	C3	C4	C5	C6
1	Good	Very good	Very good	Good	Good	Very good
2	Very good	Good	Good	Good	Good	Good
3	Good	Good	Good	Very good	Good	Good
4	Good	Good	Good	Good	Very good	Good
5	Good	Very good	Good	Good	Good	Good
6	Good	Good	Very good	Very good	Good	Good
7	Very good	Good	Good	Good	Good	Good
8	Very good	Good	Good	Good	Good	Very good
9	Very good	Very good	Good	Good	Good	Good
10	Good	Good	Very good	Good	Very good	Good

based on table 5, the next process is to normalize the criteria values based on the weights in Table 3. normalization results for all alternatives can be seen in Table 6.

Table 6. The results of the normalization of the criteria values

A	C1	C2	C3	C4	C5	C6
1	0.257	0.457	0.457	0.257	0.257	0.457
2	0.457	0.257	0.257	0.257	0.257	0.257
3	0.257	0.257	0.257	0.457	0.257	0.257
4	0.257	0.257	0.257	0.257	0.457	0.257
5	0.257	0.457	0.257	0.257	0.257	0.257
6	0.257	0.257	0.457	0.457	0.257	0.257
7	0.457	0.257	0.257	0.257	0.257	0.257
8	0.457	0.257	0.257	0.257	0.257	0.457
9	0.457	0.457	0.257	0.257	0.257	0.257
10	0.257	0.257	0.457	0.257	0.457	0.257

The values in table 6 above are obtained from the results of the initial value transformation of the criteria with the weight value of each sub-criteria calculated using ROC weighting. Then the normalized result value will be converted into a utility value using equation 2. The following utility values for each criterion and alternative are shown in table 7.

Table 7. Value of Utility

A	C1	C2	C3	C4	C5	C6
1	0,520	1	1	0,520	0,520	1
2	1	0,520	0,520	0,520	0,520	0,520
3	0,520	0,520	0,520	1	0,520	0,520
4	0,520	0,520	0,520	0,520	1	0,520
5	0,520	1	0,520	0,520	0,520	0,520
6	0,520	0,520	1	1	0,520	0,520
7	1	0,520	0,520	0,520	0,520	0,520
8	1	0,520	0,520	0,520	0,520	1
9	1	1	0,520	0,520	0,520	0,520
10	0,520	0,520	1	0,520	1	0,520

Based on the utility value generated, the next step is to determine the final value. Equation 1 is used to calculate the final amount (NA) in the Smarter method, as can be seen in table 8 to table 10 below.

Table 8. Final scores using the SMARTER method

A	C1	C2	C3	C4	C5	C6	NA
1	0,067	0,129	0,129	0,067	0,067	0,028	0,487
2	0,129	0,067	0,067	0,067	0,067	0,015	0,412
3	0,067	0,067	0,067	0,129	0,067	0,015	0,412
4	0,067	0,067	0,067	0,067	0,129	0,015	0,412
5	0,067	0,129	0,067	0,067	0,067	0,015	0,412
6	0,067	0,067	0,129	0,129	0,067	0,015	0,474
7	0,129	0,067	0,067	0,067	0,067	0,015	0,412
8	0,129	0,067	0,067	0,067	0,067	0,028	0,426
9	0,129	0,129	0,067	0,067	0,067	0,015	0,474
10	0,067	0,067	0,129	0,067	0,129	0,015	0,474

Table 9. Assessment of Job Performance with the SMARTER Method

A	C1	C2	C3	C4	C5	C6	NA	%	R
1	0,067	0,129	0,129	0,067	0,067	0,028	0,487	48.7	1
6	0,067	0,067	0,129	0,129	0,067	0,015	0,474	47.4	2
9	0,129	0,129	0,067	0,067	0,067	0,015	0,474	47.4	3
10	0,067	0,067	0,129	0,067	0,129	0,015	0,474	47.4	4
8	0,129	0,067	0,067	0,067	0,067	0,028	0,426	42.6	5
2	0,129	0,067	0,067	0,067	0,067	0,015	0,412	41.2	6
3	0,067	0,067	0,067	0,129	0,067	0,015	0,412	41.2	7
4	0,067	0,067	0,067	0,067	0,129	0,015	0,412	41.2	8
5	0,067	0,129	0,067	0,067	0,067	0,015	0,412	41.2	9
7	0,129	0,067	0,067	0,067	0,067	0,015	0,412	41.2	10

Table 10. Assessment of Job Performance with the SMARTER Method

A	C1	C2	C3	C4	C5	C6	Aktul	SMARTER
1	Good	Very good	Very good	Good	Good	Very good	Good	Enaugh
2	Very good	Good	Good	Good	Good	Good	Good	Enaugh
3	Good	Good	Good	Very good	Good	Good	Good	Medium
4	Good	Good	Good	Good	Very good	Good	Good	Medium
5	Good	Very good	Good	Good	Good	Good	Good	Enaugh
6	Good	Good	Very good	Very good	Good	Good	Good	Enaugh
7	Very good	Good	Good	Good	Good	Good	Good	Enaugh
8	Very good	Good	Good	Good	Good	Very good	Good	Enaugh
9	Very good	Very good	Good	Good	Good	Good	Good	Good
10	Good	Good	Very good	Good	Very good	Good	Good	Enaugh

Based on table 8, which shows the final value of the calculation using the SMARTER method, a ranking (R) of the highest alternative final value to the lowest alternative value can be formed as in table 9.

The results of the ranking in table 9 can be used as a test by comparing the results of the decision holders' actual decisions with the results of applying the SMARTER method in table 10.

Based on the results of the comparison table 10 with the application of the method to the actual assessment SMARTER able to provide alternative recommendations for the decision if the decision-making process based on the weighted criteria or the level of interest among different criteria.

IV. CONCLUSION

Work Behavior Assessment carried out by applying the Simple Multi-Attribute Rating Technique Exploiting Ranges (SMARTER) and Behaviorally Anchor Rating Scale (BARS) methods. The result shows that the work behaviour appraisal process requires measurable and transparent standards, is objective and produces feedback on employee work behaviour achievements.

By applying these two methods, a more objective assessment of work behaviour is produced by applying a behavioural assessment with several anchors used, as well as producing behavioural assessment feedback in the form of a final value that becomes a reference in decision making for management [17].

REFERENCES

[1] Dessler, G. (2013). Human Resource Management , 13th Edition. London: Pearson Prentice Hall Inc.

[2] Dewi, M. A., Murad, D. F., & Rosdiana. (2019). Implementation Of The Smart Models For Application Development Employee Performance Appraisal. International Conference on Sustainable Information Engineering and Technology (SIET) (pp. 364-269). Lombok, Indonesia: IEEE.

[3] Edward, W., & F.Hutton, B. (1994). Edwards, W. And Barron, F.H. Smarts And Smarter: Improved Simple Methods For Multi Attribute Utility Measurement, Organizational Behaviour And Human Decision Process, 1994. Organizational Behavior and Human Decision Processes, 306-325.

[4] Evita, S., Zusnita Muizu, W., & Atmojo, R. (2017). Employee Performance Appraisal Using Behaviorally Anchor Rating Scale Methods And Management By Objectives (Case Study at PT Qwords Company International). Pebkis Jurnal, 18-32.

[5] Han, J., & Kamber, M. (2000). Data Mining : Concepts and Techniques. San Fransisco: Morgan Kaufmann Publishers.

[6] Haryanti, D., Nasution, H., & Sukamto, A. S. (2016). Dwi Haryanti, Helfi Nasution, Anggi Srimurdianti S. Decision Support System for Admission of Bidikmisi Full Scholarship Replacement Students at Tanjungpura University by Applying the Smarter Method. Journal of Information Technology and Systems, 1-7.

[7] Hollenbeck, J., Noe, R., Wright, P., & Gerhart, B. (2012). Human Resource Management. In J. Hollenbeck, R. Noe, P. Wright, & B. Gerhart, Human Resource Management Gaining Competitive Advantage. Boston: McGraw-Hill Education.

[8] Martin-Raugh, M., Tannenbaum, R., Tocci, C., & Reese, C. (2016). Behaviorally anchored rating scales: An application for evaluating. Teaching and Teacher Education, 414-419.

[9] Morais, D., & Schramm, F. (2012). Decision Support Model for Selecting and Evaluating Suppliers in the Construction Industry. Pesquisa Operacional, 643-662.

[10] Saleh, A., Puspita, K., Sanjaya, A., Daifiria, & Giovani. (2018). Implementation of Equal Width Interval Discretization on SMARTER Method for Selecting mputer Laboratory Assistant . International Conference on Cyber and IT Service Management (CITSM 2018) (pp. 1-4). Medan: IEEE.

[11] Schwab, D., Heneman III, H., & DeCOTIIS, T. (2006). Behaviorally Anchored Rating Scales: A Review Of The Literature . Personnel Psychology, 549-562.

[12] Sevima. (2017, Juny 12). Sevima. Retrieved from Sentra Vidya Utama: https://sevima.com/pentingnya-evaluasi-kinerja-dosen/

[13] Siregar, D., Arisandi, D., Usman, A., & Irwan, D. (2017). Research Of Simple Multi-Attribute Rating Technique For Decision Support. Journal of Phisics Conference Series, 1-6.

[14] Turban, E., Aronson, J., & Liang, T.-P. (2005). DECISION support systems and intelligent systems. In D. Prabatini, sistem pendukung keputusan dan sistem cerdas. Yogyakarta: Andi Offset.

[15] Vance, R., Kuhnert, K., & Farr, J. (1978). Interview judgments: Using external criteria to compare behavioral and graphic scale ratings. Organizational Behavioral and Human Performance, 279-294.

[16] W.S, W. (1995). Psychology of Performance Appraisal. Jakarta: Gramedia.

[17] Wardhana, A., & Betrianis. (2006). Supply Chain Management Performance Assessment Using Fuzzy Set Theory Method and Typical Performance Measurement Method,. Jurnal of Technology, 221-229.

e-ISSN: 2614-8404
p-ISSN:2776-3234

Analysis of the Use of Particle Swarm Optimization on Naïve Bayes for Classification of Credit Bank Applications

Yoga Religia[1*]), Gatot Tri Pranoto[2], I Made Suwancita [3]
[1]Informatics Engineering Study Program, Faculty of Engineering, Pelita Bangsa University
[2]Information Systems Study Program, Faculty of Creative Industries and Telematics, Trilogy University
[3]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
email: [1]yoga.religia@pelitabangsa.ac.id , [2]gatot.tp@gmail.com, [3]isuwancita@gmail.com

Abstract –The selection of prospective customers who apply for credit in the banking world is a very important thing to be considered by the marketing department in order to avoid non-performing loans. The website www.kaggle.com currently provides South German Credit data in the form of supervised learning data. The use of data mining techniques makes it possible to find hidden patterns contained in large data sets, one of which is using classification modeling. This study aims to compare the classification of South German Credit data using the Naïve Bayes algorithm and compare the classification of South German Credit data using the Naïve Bayes algorithm with particle swarm optimization (PSO). The test was carried out using a confusion matrix to determine the accuracy, precision and recall values of the research model. Based on the test, it is known that PSO is able to increase the accuracy and recall of Nave Bayes, but PSO has not been able to increase the precision value of Nave Bayes. The test results show that PSO optimization gives Naïve Bayes an increase in the value of accuracy by 0.46%, and gives Naïve Bayes an increase in recall value by 3.02%.

Keywords – Data Mining, Classification, Nave Bayes, PSO Optimization, bank credit acceptance.

I. INTRODUCTION

The banking marketing department needs to select prospective customers to find out which customers can be given credit financing by considering various factors. Credit financing is the provision of funds by the bank to the customer based on a loan agreement that requires the customer to repay the loan within a certain period of time.[1]. Therefore, the selection of prospective customers is needed so that a marketing bank is able to keep their customers from experiencing non-performing loans[2]. One way that can be used to reduce the possibility of non-performing loans is to utilize data mining techniques, so that it is possible to mine information from pre-existing credit application data sets.[3].

In general, data mining is divided into two categories, namely predictive and descriptive. Predictive methods can be done with a classification model. The use of the classification model can be done by changing the data record into a set of the same class[4]. Nowsitewww.kaggle.com has provided the South German Credit data set consisting of 21 attributes with 800 instances of credit application and there is no missing value, so that it can be used to build a creditworthiness classification model [5]. The label attribute contained in the South German Credit data is the "Credit" attribute with 600 instances with the description "accepted" and 200 instances with the description "rejected", thus making South German Credit data including imbalance data.

It takes a good algorithm to create an optimal classification model. One algorithm that has been widely used for classification modeling with good performance is the Naïve Bayes algorithm. Several previous studies have stated that the Naïve Bayes algorithm is able to provide better classification performance when compared to other classification algorithms[6] [7] [8]. The Naïve Bayes algorithm can also be used on imbalanced data[9] [10], so it is suitable for classifying South German Credit data. Currently, independent assumptions are rarely discussed in the Naïve Bayes classification. One way to try independent assumptions in the Naïve Bayes algorithm is by attribute weighting[11]. It is necessary to propose an attribute weighting method to reduce the independent assumption[12]. One of the weighting optimization methods that can be used is to use particle swarm optimization (PSO).[13].

PSO has significant advantages in handling non-linear fittings and multi-input parameters [14]. PSO does not have evolution operators such as crossover and mutation, so it is easy to implement and there are very few parameters to adjust[15]. Based on several previous studies, it was stated that the combination of PSO and Naive Bayes was able to provide better imbalance data classification performance results than using Naive Bayes alone.[16] [17], even PSO is able to increase the accuracy of Naive Bayes by more than 10% [18].

Based on previous research showing that both PSO is able to improve classification performance on Nave

Bayes, it is necessary to do further testing regarding the use of PSO optimas on Naïve Bayes for South German Credit data classification. This study will compare the Naïve Bayes classification on South German Credit data with and without PSO optimization.

A. Bank Credit Financing

The word credit comes from the Italian word credere which means trust. The trust referred to here is the trust of the creditor that the debtor will return the loan and the interest in accordance with the agreement that has been agreed by both parties.[19]. The implementation of the granting of credit is carried out through several steps, namely credit application, credit application examination, credit analysis, credit approval, credit realization and the last is credit supervision.[20]. In general, most of the bank's wealth is in the form of credit which is a source of bank income which is commonly referred to as productive assets. Management must use the precautionary principle so that loans are in the current category. Often there are several customers whose interest and principal payments are not smooth which makes them fall into the category of non-performing loans (NPL). The higher the NPL indicates the greater the potential loss, so the bank must be able to reduce its lending[21].

B. Data Mining

Data mining has been around since the 1990s as an effective way to extract previously unknown patterns and information from a data set [22]. Data mining is one of the most important fields in research that aims to obtain information from data sets. Data mining is the process of extracting meaningful information and structures in complex data sets[23]. In its implementation, data mining can use various parameters to examine data including association, classification and clustering. Data mining involves key steps which include problem definition, data exploration, data preparation, modeling, and evaluating and deployment[24].

Data mining techniques are used to find relationships between data to perform classifications that predict the values of several variables (classification), or to divide known data into groups that have similar characteristics (clustering). Using data mining techniques it is possible to search, analyze, and sort through large data sets to discover new patterns, trends, and relationships contained within them.[25].

C. Classification with Naïve Bayes

The Naïve Bayes algorithm is a supervised learning algorithm based on the Bayes Theorem with the assumption of independence between predictors. That is, features in a class do not depend on other features[26]. Naïve Bayes is widely used to solve classification problems in real-world applications, this is because it is easy to build and interpret data, and has good performance.[12]. The Naive Bayes classifier can also be used for continuous and categorical variables. It is based on the Bayes formula which is the probability of event A given proof of B which can be seen in the following equation[27]:

$$\mathrm{P(A,B) = P(A)P(B)} \qquad (1)$$

Through equation (1) and using the concept of Bayes' theorem, the final equation of the Naïve Bayes algorithm is obtained as follows:

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)\,P(A)}}{\mathrm{P(B)}} \qquad (2)$$

Based on equation (2), it is known that A represents the class and B represents the instance. A is the dependent event which means the predicted variable and B is the previous event which means the predictor attribute. The last step of the Naive Bayes algorithm is to find the maximum probability that will be used as a predictor class.

D. Particle Swarm Optimization(PSO)

Particle swarm optimization or commonly abbreviated as PSO is an optimization technique whose concept is based on the behavior of a swarm of insects, such as ants, termites, or bees. [28]. The PSO approach is like a collection of particles that simultaneously explore the problem search space with the aim of finding the optimal global configuration[29]. PSO has proven to be very effective in solving various engineering problems and solving them very quickly[30].

The basic assumption behind the PSO algorithm is that birds find food in groups and not individually. This gives rise to the assumption that information is shared in flocking. The herd initially has a population of random solutions. Each potential solution is called a particle (agent), is assigned a random velocity and is flown through the problem space[30]. All particles have a memory and each particle keeps track of the previous best position (Pbest) and the corresponding match value. Flock has another value called Gbest, which is the best value of all Pbests. By using this concept, PSO can provide a technique for solving attribute selection quickly.

II. RESEARCH METHODOLOGY

A. Data used

This study uses secondary data in the form of a South German Credit data set taken from the sitewww.kaggle.com [5]. The number of data instances contained in the South German Credit data is 800 instances consisting of 21 attributes and there is no missing value, so there is no need for pre-processing data. Based on the existing 21 attributes, there is 1 label attribute contained in the South German Credit data, namely the "Credit" attribute. On the label there are 600 instances with the description "good" and 200 instances with the description "bad", so that the South German Credit data includes imbalance data.

South German Credit Data chosen because it is free from missing values, so there is no need for preprocessing data to be used in making classification models. As for more clearly about 21 attributes and 1 label from South German Credit data, it can be seen in Table 1.

Table1. Data AttributeSouth German Credit

Attribute	Information
status	status of the debtor's checking account with the bank (categorical)
duration	credit duration in months (quantitative)
credit history	history of compliance with previous or concurrent credit contracts (categorical)
purpose	purpose for which the credit is needed (categorical)
amount	credit amount in DM (quantitative; result of monotonic transformation; actual data and type of trans...
savings	debtor's savings (categorical)
employmen t duration	duration of debtor's employment with current employer (ordinal; discretized quantitative)
installment rate	credit installments as a percentage of debtor's disposable income (ordinal; discretized quantitative...
personal status sex	combined information on sex and marital status; categorical; sex cannot be recovered from the variab...
other debtors	Is there another debtor or a guarantor for the credit? (categorical)
present residence	length of time (in years) the debtor lives in the present residence (ordinal; discretized quantitative...
property	the debtor's most valuable property, ie the highest possible code is used. Code 2 is used, if code...
age	age in years (quantitative)
other installment plans	installment plans from providers other than the credit-giving bank (categorical)
housing	type of housing the debtor lives in (categorical)
number credits	number of credits including the current one the debtor has (or had) at this bank (ordinal, discretiz...
job	quality of debtor's job (ordinal)
people liable	number of persons who financially depend on the debtor (ie, are entitled to maintenance) (binary,d...
telephone	Is there a telephone landline registered on the debtor's name? (binary; remember that the data are f...
foreign workers	Is the debtor a foreign worker? (binary)
credit risk	Has the credit contract been complied with (good) or not (bad) ? (binary)

B. Research Model

The classification model built in this study was carried out using South German Credit data. The label used is the attribute "Credit Risk" with the values "Good" and "Bad". As many as 77% of the instances in the South German Credit data are instances with the class label "good", while the rest are instances with the label "bad". Tests in this study were carried out 2 times which will later be analyzed the results obtained. The first test was carried out using Naïve Bayes with PSO optimization, while the second test was carried out using Nave Bayes without PSO optimization.

Spit validation used as a process of validating research data which aims to divide South German Credit data into training data and testing data. Split validation is used to divide South German Credit data into training and testing data with a comparison of 90% and testing data of 10%. The training data will be used for classification modeling using the Naïve Bayes algorithm. The resulting model is then used as an apply model for use in data testing. After the classification has been carried out, the performance of the formed classification model is measured in the form of accuracy, precision, and recall values.

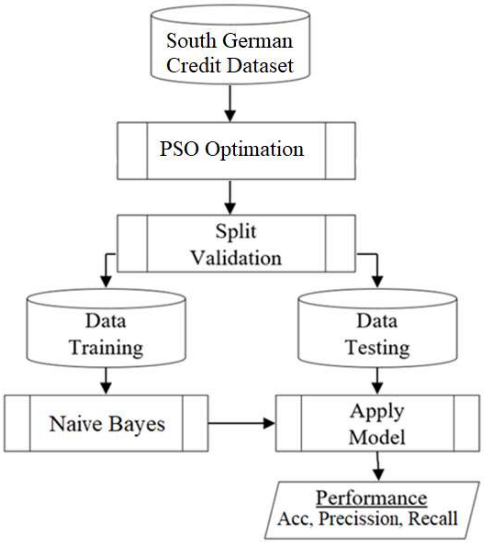


Figure 1. First Test: Nave Bayes Classification with PSO Optimization

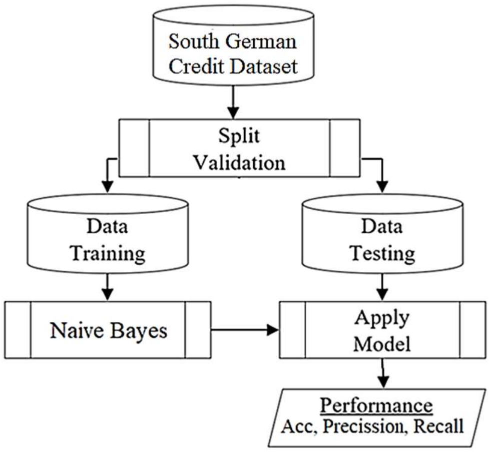


Figure 2. Second test: Nave Bayes Classification without PSO Optimization

Based on Figure 1 and Figure 2 it shows that the testing in this study was carried out 2 times, namely: (1) Classification of South German Credit data using Naïve Bayes with PSO optimization, (2) Classification of South German Credit data using Naïve Bayes without PSO optimization. The performance results of the two tests will be compared and then analyzed to obtain research findings.

III. RESULTS AND DISCUSSION

A. Testing Step

The testing tools in this study used RapidMiner version 5.0. The use of RapidMiner tools is done because RapidMiner can be used for rapid prototyping, and supports all steps of the data mining process[31]. The first step in making this research model is to call South German Credit data. The second step is to perform the multiply function to perform two tests at once, namely testing using PSO optimization and testing without using PSO optimization. The third step is to distribute the data into the split validation process. Validation process by dividing training data by 90% and testing data by 10%

from South German Credit data. More clearly about the data calling and validation process in this study can be seen in Figure 3.

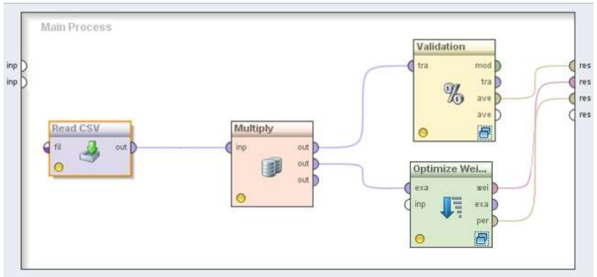


Figure 3. Data Calling and Validation Process

In each validation process in Figure 3, it contains a learning process using the Naïve Bayes algorithm which is then applied to the model to measure its accuracy, precision and recall performance. The learning process formed in this study can be seen in Figure 4.

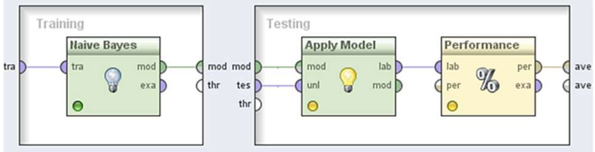


Figure 4. Learning Process and Apply Model

After the entire research model has been formed, the last step is to run the model that has been built in RapidMiner to see the results of its accuracy, precision and recall.

B. Test result

After 2 tests, the accuracy, precision, and recall values of the two models were obtained. More complete test results can be seen in Table 2.

Table2. Test result

No	Algorithm	Accuracy	Precision	Recall
1	Naive Bayes	85.43%	89.37%	85.80%
2	PSO + Naive Bayes	85.89%	88.23%	88.82%

Based on Table 2, it can be seen that PSO is able to increase the accuracy and recall of Nave Bayes, but PSO has not been able to increase the precision value of Nave Bayes. The test results show that with an accuracy of 85.89%, PSO optimization gives Naïve Bayes an increase in accuracy value of 0.46% and an increase in recall value of 3.02% for South German Credit data classification. However, it turns out that the use of PSO also reduces the precision value by 1.14% from the use of Naïve Bayes for South German Credit data classification. This is presumably because of the 20 attributes (non attribute labels) in the South German Credit data, it turns out that there are only 5 attributes that are weighted by PSO. This weighting result also explains why the increase in accuracy and recall provided by PSO is not too large.

In Table 3 it can be seen that there are only 5 attributes that are weighted by PSO. This shows that, based on PSO optimization, these 5 attributes are the most important to consider when classifying South German Credit data. The attributes are: credit history, savings, property, age, and job.

Table 3. PSO Weighting Results on DataSouth German Credit

Attribute	Weighting
status	0
duration	0
credit history	1
purpose	0
amount	0
savings	1
employment duration	0
installment rate	0
personal status sex	0
other debtors	0
present residence	0
property	1
age	1
other installment plans	0
housing	0
number credits	0
job	1
people liable	0
telephone	0
foreign workers	0

C. Results Discussion

Based on the results of the tests that have been carried out, it is known that the use of PSO optimization on Nave Bayes for the classification ofSouth German Credit data is able to improve the performance of Naïve Bayes in terms of accuracy and recall even though it is not too big. The small increase in performance given by PSO is thought to be because the attributes that are weighted by PSO are only 5 attributes out of 20 predictor attributes in South German Credit data. This makes the probability calculation process on Naïve Bayes become irrelevant. Even looking at the precision side, it turns out that the use of PSO optimization actually makes Naïve Bayes' precision performance decrease.

Although the optimization of PSO does not give maximum results, by using PSO it can be seen which attributes can be used as evaluation priorities to consider loan application approval. By looking at the attributes given the weighting by PSO, it can be used as a reference to consider these attributes as the main focus to avoid the risk of non-performing loans. The attributes that are given weighting by PSO are: credit history, savings, property, age, and job. This finding is expected to provide a practical contribution to the decision to provide credit by marketing parties in order to minimize the occurrence of non-performing loans.

IV. Conclusion

This research has tested the use of Naïve Bayes algorithm and PSO optimization to classify South German Credit data. Based on the tests that have been carried out, some results are shown as follows:

1. PSO optimization is able to improve the performance of Naïve Bayes in classifying South German Credit data in terms of accuracy and recall, although it does not have a big impact.
2. This study found that the use of PSO has not been able to improve the performance of Naïve Bayes in classifying South German Credit data in terms of precision, even the precision performance of Nave Bayes has decreased in value.

3. By using PSO optimization, it can be seen that there are 5 attributes that need to be the main consideration in lending in the banking world, namely credit history, savings, property, age, and job.

This study has not been able to provide a good enough accuracy value for the classification of South German Credit data, so further research is needed to obtain a better classification model. Based on the findings of this study, it is suggested that further research can apply different optimization methods such as Bagging, Genetic Algorithm, Adaboost or other optimizations to significantly improve the performance of Naïve Bayes in classifying South German Credit data.

REFERENCES

[1] AT Rahmawati, M. Saifi and RR Hidayat, "Analysis of Credit Provision Decisions in Minimizing Non-Performing Loans," Journal of Business Administration, vol. 35, no. 1, pp. 179-186, 2016.

[2] S. Somadiyono and T. Tresya, "Criminal Responsibility for Marketing according to the Banking Law for Non-performing Financing at Bank Muamalat Indonesia, Tbk," Journal of Lex Specialis, vol. 21, pp. 22-38, 2015.

[3] S. Masripah, “Comparison of Data Mining Classification Algorithms for Evaluation of Credit Provisions,” Bina Insani ICT Journal, vol. 3, no. 1, pp. 187-193, 2016.

[4] S. Umadevi and KSJ Marseline, “A Survey on Data Mining Classification Algorithms,” at the International Conference on Signal Processing and Communication, Coimbatore, India, 2017.

[5] “Kaggle,” kaggle.com, 2020. [Online]. Available: https://www.kaggle.com/c/south-german-credit-prediction/overview/data-overview. [Accessed 2 November 2020].

[6] IAA Amra and AYA Maghari, “Students Performance Prediction Using KNN and Naïve Bayesian,” at the 8th International Conference on Information Technology (ICIT), Al-Zaytoonah University of Jordan, Jordan, 2017.

[7] F. Osisanwo, J. Akinsola, O. Awodele, JO Hinmikaiye, O. Olakanmi and J. Akinjobi, “Supervised Machine Learning Algorithms: Classification and Comparison,” International Journal of Computer Trends and Technology (IJCTT), vol. 48, no. 3, pp. 128-138, 2017.

[8] EN Azizah, U. Pujianto, E. Nugraha and Darusalam, “Comparative Performance Between C4.5 and Naive Bayes Classifiers in Predicting Student Academic Performance in A Virtual Learning Environment,” in the 4th International Conference on Education and Technology (ICET), Malang, Indonesia, 2018.

[9] K. Madasamy and M. Ramaswami, “Data Imbalance and Classifiers: Impact and Solutions from A Big Data Perspective,” International Journal of Computational Intelligence Research, vol. 13, no. 9, pp. 2267-2281, 2017.

[10] EM Hassib, AI El-Desouky, E.-SM El-Kenawy and SM El-Ghamrawy, “An Imbalanced Big Data Mining Framework for Improving Optimization Algorithms Performance,” Journal & Magazines, vol. 7, no. 1, pp. 170774-170795, 2019.

[11] S. Chen, GI Webb, L. Liu and X. Ma, “A Novel Selective Naïve Bayes Algorithm,” Knowledge-Based Systems, vol. 192, pp. 1-15, 2020.

[12] L. Jiang, L. Zhang, L. Yu and D. Wang, “Class-Specific Attribute Weighted Naive Bayes,” Pattern Recognition, vol. 88, no. 1, pp. 321-330, 2019.

[13] S. Ernawati, R. Wati, N. Nuris, LS Marita and ER Yulia, “Comparison of Naïve Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application,” Journal of Physics: Conference Series , vol. 1641, pp. 1-7, 2020.

[14] X. Liu, Z. Liu, Z. Liang, S.-P. Zu, JAFO Correia and AMPD Jesus, “PSO-BP Neural Network-Based Strain Prediction of Wind Turbine Blades,” Materials, vol. 12, no. 12, pp. 2-15, 2019.

[15] S. Srivastava, J. Gupta and M. Gupta, “PSO & Neural-Network Based Signature Recognition for Harmonic Source Identification,” in IEEE Region 10 International Conference TENCON, Singapore, 2009.

[16] M. Misdram, E. Noersasongko, A. Syukur, Purwanto, M. Muljono, HA Santoso and DRIM Setiadi, "Analysis of Imputation Methods of Small and Unbalanced Datasets in Classifications using Naïve Bayes and Particle Swarm Optimization," in International Seminar on Application for Technology of Information and Communication (ISemantic), Semarang, Indonesia, 2020.

[17] I. Romli, T. Pardamean, S. Butsianto, TN Wiyatno and EB Mohamad, “Naive Bayes Algorithm Implementation Based on Particle Swarm Optimization in Analyzing the Defect Product,” Journal of Physics: Conference Series, vol. 1845, no. 1, pp. 1-6, 2021.

[18] J. Li, L. Ding and B. Li, “A Novel Naive Bayes Classification Algorithm Based on Particle Swarm Optimization,” The Open Automation and Control Systems Journal, vol. 6, no. 1, pp. 747-753, 2014.

[19] MS Hasibuan, Banking Fundamentals, Jakarta: PT Bumi Aksara, 2004.

[20] R. Widayati and M. Efriani, “Activities of Business Loans at PT. Batang Kapas Rural Bank," in OSF Preprints, Batang, Indonesia, 2019.

[21] B. Panuntun and Sutrisno, "Determining Factors in Banking Credit Disbursement Case Study in Conventional Banks in Indonesia," Dewantar Journal of Accounting & Finance Research, vol. 1, no. 2, pp. 57-66, 2018.

[22] D. Tomar and S. Agarwal, “A survey on Data Mining approaches for Healthcare,” International Journal of Bio-Science and Bio-Technology, vol. 5, no. 5, pp. 241-266, 2013.

[23] MS Başarslan and ID Argun, “Classification Of A Data Bank Set On Various Data Mining Platforms,” in Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, Turkey, 2018.

[24] V. Krishnaiah, G. Narsimha and N. Chandra, “Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques,” International Journal of Computer Science and Information Technologies, vol. 4, no. 1, pp. 39-45, 2013.

[25] K. Sumiran, “An Overview of Data Mining Techniques and Their Application in Industrial Engineering,” Asian Journal of Applied Science and Technology, vol. 2, no. 2, pp. 947-953, 2018.

[26] K. Yadav and R. Thareja, “Comparing the Performance of Naive Bayes and Decision Tree Classification Using R,” IJ Intelligent Systems and Applications, vol. 12, pp. 11-19, 2019.

[27] S. Sankaranarayanan and TP Perumal, “Analysis of Naïve Bayes Classification for Diabetes Mellitus,” International Journal of Computer Sciences and Engineering, vol. 6, no. 12, pp. 520-524, 2018.

[28] M. Jaenal, A. Nugroho and I. Romli, “Analysis of Effectiveness Particle Swarm Optimization in Improving The Performance of Naïve Bayes Algorithm,” in ICID Proceedings, Yogyakarta, Indonesia, 2018.

[29] B. Chopard and M. Tomassini, “Particle Swarm Optimization,” in An Introduction to Metaheuristics for Optimization, Springer, Cham, Natural Computing Series, 2018, p. 97–102.

[30] YS Haruna, YA Yisah, GA Bakare, MS Haruna and SO Oodo, “Optimal Economic Load Dispatch of the Nigerian Thermal Power Stations Using Particle Swarm Optimization (PSO),” The International Journal Of Engineering And Science (IJES), vol. 6, no. 1, pp. 17-23, 2017.

[31] A. Jeyaraj, R. S and MR Raja, “A study of classification algorithms using Rapidminer,” International Journal of Pure and Applied Mathematics, vol. 119, no. 12, pp. 15977-15988, 2018.

e-ISSN: 2614-8404
p-ISSN:2776-3234

Water Quality Monitoring System with Parameter of pH, Temperature, Turbidity, and Salinity Based on Internet of Things

Yazi Adityas[1*)], Sasmitoh Rahmad Riady[2], Muchromi Ahmad [3], Moh Khamim [4], Khalis Sofi [5*)],
[1]Direktorat Teknologi Informasi, Universitas Pelita Bangsa
[2]Magister Science in Information Technology, Fakultas Computing, President University
[3,4,5]Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
Email: [1]yaziadityas@pelitabangsa.ac.id, [2]sasmitohrr@student.president.ac.id,[3]muchromiahmad@gmail.com,
[4]moh.khamim04@gmail.com, [5]khalissofi@mhs.pelitabangsa.ac.id,

Abstract – This research aims to monitor the quality of water used for aquariums. The physical parameters used are water pH, water temperature, water turbidity, and water salinity. Using a pH sensor, temperature sensor, turbidity sensor, and salinity conductivity sensor with Arduino as the controller. The prototype method used in this research, starting from the formulation, research, building stages to testing and evaluating the results of the research. The working process of the system is when the system is activated, the sensors will detect and capture the amount of value contained in the water, then the data from the sensor is sent to a database in the cloud using an ethernet shield that is connected to the media router as a liaison for the internet network then displayed on the website dashboard in the form of graphs and monitoring record tables in real time. The sensors function to detect water quality, where quality standards have been set in this system, namely temperature standards of 27-30°C, pH standards of 7.0-8.0, turbidity standards of 2.5-5 ntu, and salinity of 20-28 ppt. If the sensor detects non-compliance with water quality standards, the buzzer in this system will sound. From the results of system testing, sensors can detect water quality in real time within 5-10 seconds. Based on the research results, this water quality monitoring system is effective to help ensure the quality of the water in the aquarium so that it always meets the standards.

Keywords – Water Quality, Internet of Things, Ph, Temperature, Turbidity, Salinity

I. INTRODUCTION

Water is the main source of need for human survival. Clean water has a very important role in improving environmental health, and plays a role in improving standards or life quality [1]. In daily life, the use of water is used in various fields such as plantations, fisheries, industry, animal husbandry, and other fields. In this pandemic period, many people use water in the field of fisheries, one of which is by keeping ornamental fish in the aquarium to fill activities while at home. [2]. To keep fish in an aquarium, of course you have to pay attention to aspects of the feasibility of clean water, such as temperature, pH, turbidity, and water salinity. Standard of pH for clean water is 7.0-8.0, temperature 27-30°C, turbidity tolerance 2.5-5 ntu, and salinity 20-28 ppt [3]. Thus, to ensure that water quality is always in a standard state, a monitoring system is needed [4] by utilizing Internet of Things technology (IoT) so that water quality can be ensured always in accordance with the standard [5]. IoT can be interpreted as all objects that can communicate with other objects [6]. Therefore, in this study, a prototype of an IoT-based water quality monitoring system will be made [7] by using the temperature sensor model DS18B20 [8], pH sensor [9], salinity sensor [10], and turbidity sensor[11]. Then the controller used is ethernet shield and Arduino [12], as well as a buzzer as an alarm if the sensor detects a non-compliance with AI quality standards [13]. The data from the sensor is sent to the database and displayed on the website dashboard [13] in real time [14].

Research on water quality monitoring has been carried out by Muhammad Faisal, et al to monitor water turbidity using the TSD-10 sensor. From this research, it can be seen that the sensor sensitivity value from the results of the TSD-10 sensor characteristics is 2 mV/NTU and the average accuracy of the measurement has a value of 93.49%. The maximum relative error of measurement is 24.64% [15]. Monitoring of water turbidity was also carried out by Muhammad Kautsar, et al by examining the turbidity level of PDAM water. After conducting calibration trials with conventional water volume measurements by accommodating the volume of water within a certain period of time in a measuring cup and getting a fairly good accuracy result, namely 98.8% [16].

Research on water temperature monitoring conducted by Jamal Maulana Hudin, et al uses NodeMcu, DS18B20 temperature sensor, and the Cayenne application as an IoT platform. The monitoring system designed can provide information on water temperature conditions in real time. The control system will turn on automatically when the temperature is outside the normal range. In the application of the system if the pool temperature is below 25°C it will display a "Bahaya DINGIN" notification and if the pool temperature is above 30°C it will display a "Bahaya PANAS" notification [17]. Arif Indra Irawan, et al conducted research on the temperature of fish ponds with the aim of improving the performance of the DS18B20 temperature sensor. The results showed that the measurement accuracy can be improved by using the linear regression method. The linear regression method in experimental measurements at temperatures of ± 3°C to ± 50°C can increase accuracy by 0.42%, RMSE by 34.4%, and increase sensor response time by about 12% -30%. The use of a normal distributed measurement rate and linear regression in calibration can increase the response time by 12%-19% but reduce the level of accuracy when compared to the linear regression method alone. The use of the linear regression method and a normal distributed measurement rate can increase the accuracy of the actual pool temperature measurement, both in the afternoon and in the morning by 90%.[18].

Yuri Rahmanto, et al monitored the pH of aquaponic water by using the Arduino Uno microcontroller. The result of the research is that by looking at the results of the water pH sensor readings, farmers can determine that the water is in good condition or not for mustard plants. The results of the water pH sensor readings have a difference that is not so far from the pH meter, which is 5.5 to 6.5 [19]. Dista Yoel Tadeus, et al also conducted research on the turbidity and pH of the water, with the object of the research being a freshwater aquarium by utilizing Internet of Things technology. Monitoring data is used to activate the actuator in the form of a water filter. The filter will be active if the water turbidity level has exceeded the specified turbidity limit. The turbidity test of the aquarium water shows that when the turbidity reaches 3000 ntu at 14.12 the pump is active and the filter works until the turbidity is at a value of 498 ntu at 17.00 and the pump turns off automatically. The pH value and water turbidity were successfully displayed in the Blynk application on the cellphone. The test results conclude that the monitoring system developed has been successfully implemented [20].

Dynar A. Wibisono, et al conducted a study with the aim of designing a monitoring system based on the internet of things and an automatic control system using a salinity sensor to monitor salt levels, a DS18B20 temperature sensor to monitor temperature, and a pH sensor SEN0161 to monitor water pH. The sensor data is processed by the Arduino Nano microcontroller and the Wi-Fi-based Wemos D1 mini board from the ESP8266 family, sending data to the firebase realtime database, then the user will monitor the salt content, temperature and pH content on the web. Tests are carried out by validating and reading sensors as well as sending and recording data on the web and controlling actuators to maintain water quality. the system is able to increase the temperature by 1°C in 1.25 minutes with an error value of ±1°C, the system is able to increase the salinity content by 2 ppt in 2 minutes with an error value of ±1 ppt, pH control can also be done by increasing the pH value by 1 which has an error of±0.7. The results show that sensor data can be sent in real time on the website at a speed of 484.75 ms using the HSDPA network and 75 ms using the LTE network [21].

II. RESEARCH METHODOLOGY

A. Prototype Method

In making this research, the researcher divides into 4 stages, namely formulation stage, research stage, building stage, testing stage and evaluation stage.

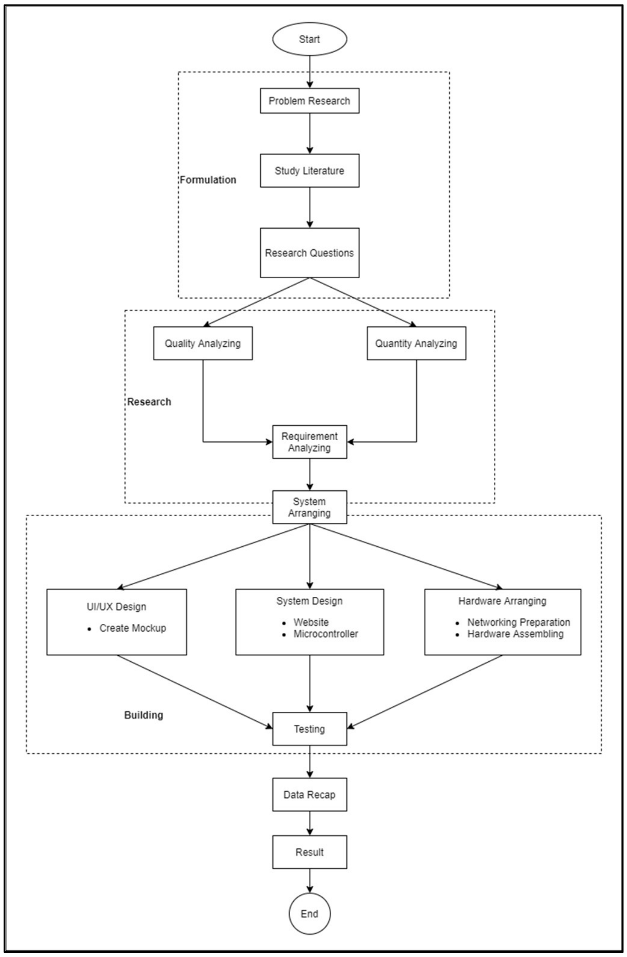


Figure. 1 Prototype Method

The following is an explanation of the methodological stages in Figure. 1.

1. Formulation, which is the management stage of the research strategy to be carried out, consisting of:
   a. Problem research, we look for the problem of what object will be researched, especially in the IT field.
   b. Study Literature, after the object of the problem is found, the researcher examines previous studies related to the object to be studied.
   c. Research Question, we describe questions related to the object of research.
2. Research, is the stage to find the criteria to be researched.
   a. Quality analysis, aims to determine the standard quality of the object of research to be carried out.
   b. Quantity analysis, aims to determine the number of needs of the object of research to be carried out.
   c. Requirement analysis, after the two criteria have been determined, a needs analysis will be carried out by the author to conduct research.
3. Building, is the execution stage of implementing the research to be carried out, starting from system design which is divided into 3 designs, including:
   a. UI/UX design, is a process to provide an initial view for users in the form of making mockups so that they can be easily used as they should be.
   b. System design or program design, is the process of

programming software in the system, namely the web system and programming on the microcontroller.

c. Hardware Arranging, is a hardware design process in the form of internet network configuration and a series of microcontroller tools.

4. Trial and Evaluation, is the final stage of research to test and evaluate research that has been designed in the previous stage which consists of:
   a. Trial, is a system testing process from research that has been previously designed.
   b. Data recap, conducting the process of collecting data on the evaluation results of the research object trial.

B. *Hardware Requirement*

The hardware devices needed in making a prototype water quality monitoring system are as follows:

Table 1. Hardware Requirement

Figure	Name	Qty	Used For
	Arduino IDE	1 pc	As a microcontroller
	Temperature Sensor model DS18B20	1 pc	Water temperature sensor with units of degrees Celsius
	pH Sensor	1 pc	Water PH level detection sensor
	Turbidity Sensor	1 pc	Water turbidity detection sensor with units of NTU
	Salinity Sensor	1 pc	Water salinity detection sensor with units of part per million ppt
	Buzzer	1 pc	As a notification in the form of a "beep" sound when the sensor detects non-standard water quality
	Ethernet Shield	1 pc	Additional port for connecting with router
	Project Board	1 pc	The place to assemble and combine sendor with arduino
	Mikrotik RB951UI	1 pc	Sebagai pengaturan dan jaringan IP arduino
	USB Arduino Cable	1 pc	As a liaison Arduino with laptop
	Jumper Cable	1 set	As a liaison sensor with arudini
	UTP/LAN Cable	1 pc	As a liaison Arduino with mikrotik

C. *Flowchart System*

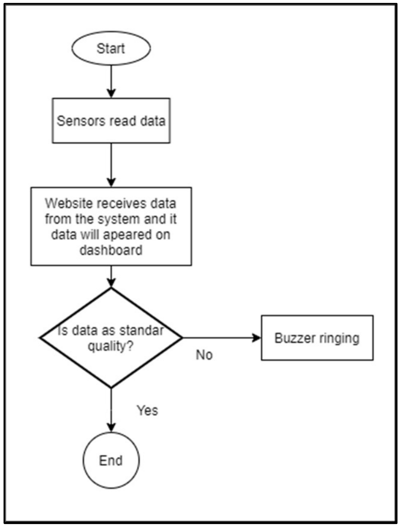

Start
Sensors read data
Website receives data from the system and it data will apeared on dashboard
Is data as standar quality?
No
Buzzer ringing
Yes
End

Figure. 2 Flowchart System

The following is an explanation of the flowchart in the figure. 2:

1. The sensors detect the water quality data in the aquarium.
2. 2. The system sends sensor data to the web server and displays it.
3. If the sensor detects a water quality discrepancy, the buzzer sounds.

D. *Diagram System*

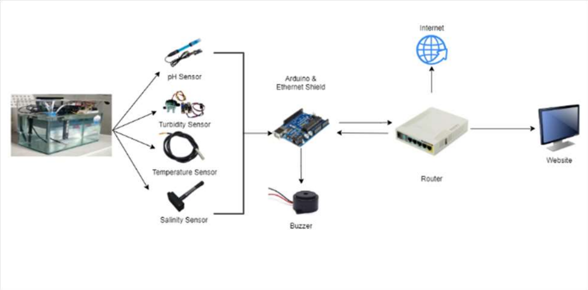


Figure. 3 Diagram System

Figures. 3 can be explained as follows.

1. In the aquarium, sensors are installed to detect water quality, namely temperature, pH, turbidity, and salt

levels. Its function is to ensure that water quality is always in accordance with standards.
2. Arduino and ethernet shield as microcontrollers that control sensors and servers that receive data from Arduino, then store and send it to the web server.
3. Buzzer sounds when the sensor detects water quality that is not up to standard

III. Results And Discussion

The results of the development of a water quality monitoring system are able to provide convenience in detecting water quality according to standards. In addition, it can also open up public insight regarding water detection in accordance with standards carried out by using automated technology. In this study, the system developed is based on Arduino which uses temperature, pH, salinity, turbidity sensors to detect water conditions using the method used, namely prototyping. The following are the results of water quality monitoring in this study.

A. Dashboard Website Monitoring Water Quality

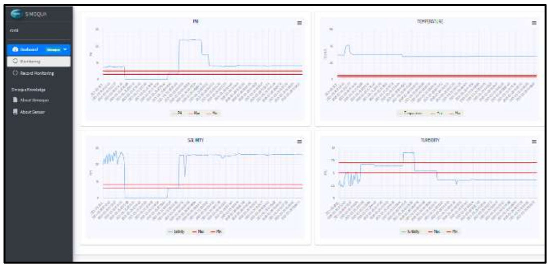


Figure. 4 Dashboard Website

Displays data graphs from sensor records that have been stored by the water quality monitoring system database. Consisting of graphs of pH, temperature, salinity, and turbidity, this website is very user friendly, so it can be set according to the needs of the type of water to be monitored. In this system, temperature standard of 27-30°C, pH standard of 7.0-8.0, turbidity standard of 2.5-5 ntu, and salinity of 20-28 ppt.

B. Graph of pH Monitoring

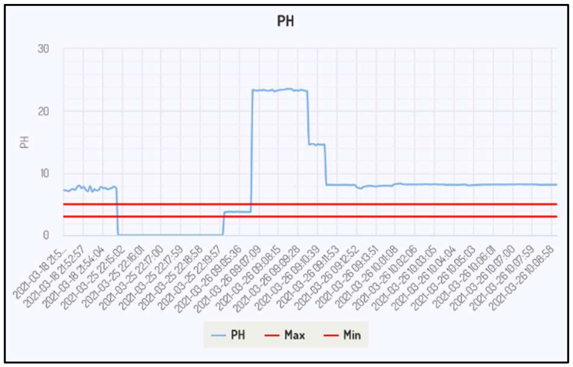


Figure. 5 Graph of pH Monitoring

On the graph of the pH data, the two red lines are the standard water pH parameters. The line below is the minimum value and the line above is the maximum value. Graphs display data in real time. Graphics can also be displayed full screen or can be downloaded with the formats that are available.

C. Graph of Temperature Monitoring

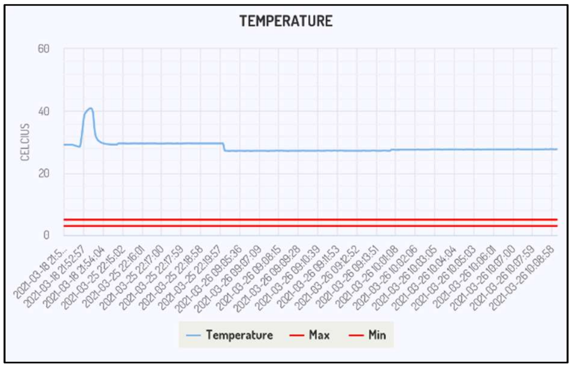


Figure. 6 Graph of Temperature Monitoring

In the temperature data graph, the two red lines are the standard water temperature parameters. The line below is the minimum value and the line above is the maximum value. Graphs display data in real time. Graphics can also be displayed full screen or can be downloaded with the formats that are available.

D. Graph of Salinity Monitoring

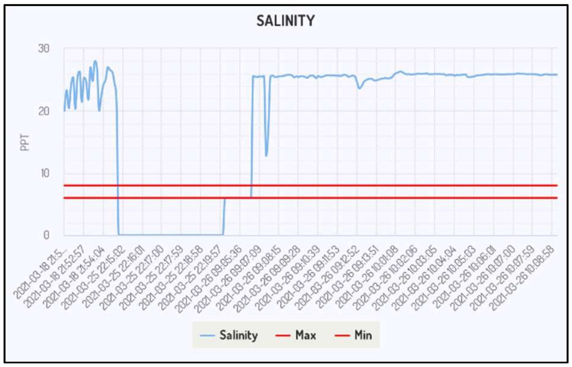


Figure. 7 Graph of Salinity Monitoring

On the salinity data graph, the two red lines are the standard water salinity parameters. The line below is the minimum value, and the line above is the maximum value. Graphs display data in real time. Graphics can also be displayed full screen or can be downloaded with the formats that are available.

E. Graph of Turbidity Monitoring

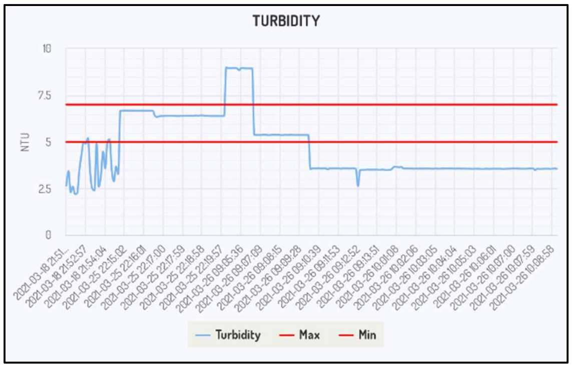


Figure. 8 Graph of Turbidity Monitoring

In the turbidity data graph, the two red lines are the standard parameters for water turbidity. The line below is the minimum value, and the line above is the maximum value. Graphs display data in real time. Graphics can also be displayed full screen or can be downloaded with the formats that are available.

F. Data Record Monitoring

This Monitoring Record contains recorded data from the sensors of the water quality system. There are data on pH, temperature, turbidity, salinity, and the date and time the data was recorded. Monitoring data can also be downloaded, and this is one of the advantages of the water quality system data report, namely that the report data attributes can be adjusted according to needs. The following is the data from the monitoring conducted on March 18, 2021.

Table 2. Monitoring Record

Temperature	pH	Turbidity	Salinity	Date & Time
29,12	7,25	2,65	20	2021-03-18 21:51:57
29,12	7,2	3,43	23,32	2021-03-18 21:52:03
29,12	7,1	2,3	20,4	2021-03-18 21:52:10
29,12	7,3	2,6	24	2021-03-18 21:52:17
29,06	7,45	2,2	25,4	2021-03-18 21:52:23
28,81	7,3	2,25	20,3	2021-03-18 21:52:30
28,62	7,77	3,48	25,4	2021-03-18 21:52:37
28,5	8	4,29	26,3	2021-03-18 21:52:43
32,31	7,6	5	21,4	2021-03-18 21:52:50
38,06	7,8	4,9	25,3	2021-03-18 21:52:57
39,69	7,3	5,2	24,7	2021-03-18 21:53:03
40,44	7,1	3,39	21,78	2021-03-18 21:53:10
40,88	7,9	2,5	27	2021-03-18 21:53:17
39,94	7	2,4	24,79	2021-03-18 21:53:23
33,25	7,4	4,9	28	2021-03-18 21:53:37
31,06	7,2	2,6	26,78	2021-03-18 21:53:44
30,25	7,3	3,2	20	2021-03-18 21:53:51
29,81	7,8	4,48	22,49	2021-03-18 21:53:57
29,56	7,6	3,59	24,23	2021-03-18 21:54:04
29,37	7,6	4,98	25	2021-03-18 21:54:11

G. Experimental Result

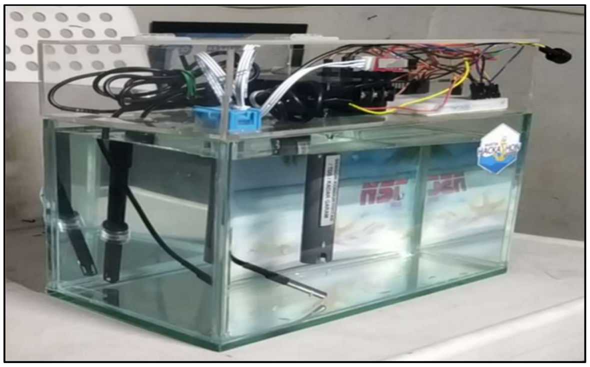


Figure. 9 Experimental Result

The prototype of the water quality monitoring system consists of a series of described hardware that is connected to a website-based system that is useful for displaying monitoring results.

IV. CONCLUSION

Based on the results of the research that has been done, it can be concluded that a water quality monitoring system based on the Internet of Things using a temperature sensor, pH sensor, turbidity sensor, and salinity sensor can monitor water in terms of temperature, pH, turbidity, and salt content whose data is displayed. on the website dashboard in real time every 5-10 seconds. The buzzer gives an audible warning when one of the sensors detects a water quality discrepancy. This can help ensure water quality is always up to standard.

REFERENCES

[1] A. Kusnandar, “Rancang Sistem Monitoring Air Layak Konsumsi Menggunakan Metode Fuzzy Tsukamoto Berbasis Android,” pp. 1–8, 2019.

[2] A. Pebrianto, R. Haryanto, and A. Pratomo, “Diseminasi Sistem Aquaponik Sebagai Salah Satu Solusi Ketahanan Pangan Di Masa Pandemi Covid-19,” *PRO Sejah. (Prosiding ...*, vol. 3, pp. 1–6, 2021.

[3] A. H. Kadafi and U. B. Luhur, “Perancangan pengendalian suhu dan salinitas air pada aquarium ikan botana biru,” vol. 3, no. 1, pp. 62–68.

[4] S. R. Riady, D. Maulana, A. Suwarno, and A. Nugroho, “Implementasi Sistem Monitoring Suhu Pada Produk Makanan di Mesin Sterilisasi Menggunakan Fuzzy Logic Berbasis Internet of Things,” *InComTech J. Telekomun. dan Komput.*, vol. 8, no. 2, pp. 121–132, 2018.

[5] U. Ulumuddin, M. Sudrajat, T. D. Rachmildha, N. Ismail, and E. A. Z. Hamidi, “Prototipe Sistem Monitoring Air Pada Tangki Berbasis Internet of Things Menggunakan Nodemcu Esp8266 Sensor dan Ultrasonik,” *Semin. Nas. Tek. Elektro 2017*, no. 2016, pp. 100–105, 2017.

[6] E. B. Lewi, U. Sunarya, and D. N. Ramadhan, “Sistem Monitoring Ketinggian Air Berbasis Internet of Things Menggunakan Google Firebase,” *Univ. Telkom, D3 Tek. Telekomun.*, vol. 1, no. 1, pp. 1–8, 2017.

[7] R. K. Putra Asmara, “Rancang Bangun Alat

Monitoring Dan Penanganan Kualitas Ait Pada Akuarium Ikan Hias Berbasis Internet Of Things (IOT)," *J. Tek. Elektro dan Komput. TRIAC*, vol. 7, no. 2, pp. 69–74, 2020.

[8] R. Arifuddin and Y. Sinatra, "Identifikasi Sensor Suhu pada Setup Awal Untuk Pengukuran Suhu Bawah Permukaan," *JEECAE (Journal Electr. Electron. Control. Automot. Eng.*, vol. 3, no. 2, pp. 209–212, 2018.

[9] M. S. Islam, E. A. Suhendi, S. Si, and M. Si, "RANCANG BANGUN REALTIME MONITORING TINGKAT KEASAMAN (PH) DAN KONDUKTIVITAS ELEKTRIK (EC) BERBASIS INTERNET OF THINGS (IOT) PADA SUNGAI CITARUM DESIGN OF REALTIME MONITORING ACID LEVEL (PH) AND ELECTRICAL CONDUCTIVITY (EC) BASED ON INTERNET OF THINGS (IOT) IN CITARUM RIVER," vol. 8, no. 2, pp. 1899–1904, 2021.

[10] S. Sukarni *et al.*, "Kontrol Kualitas Air Kolam Ikan Lele Berbasis Microbubbles dan Internet of Things (IOT)," *Pros. Has. Pengabdi. Kpd. Masyakat*, no. Hapemas 2, pp. 224–234, 2018.

[11] H. R. Iskandar, D. I. Saputra, and H. Yuliana, "Eksperimental Uji Kekeruhan Air Berbasis Internet of Things Menggunakan Sensor DFRobot SEN0189 dan MQTT Cloud Server," *J. Umj*, no. Sigdel 2017, pp. 1–9, 2019.

[12] I. Fauzi, S. Komputer, F. T. Informasi, U. B. Luhur, P. Utara, and K. Lama, "Monitoring Ketinggian dan Suhu Air Dalam Tangki Berbasis Web Menggunakan Arduino Uno & Ethernet Shield," *Bit*, vol. 14, no. 1, pp. 39–44, 2017.

[13] A. Meifriyan Pratama, D. Meidelfi, and D. Prayama, "RETRACTED: Monitoring Air Pada Water Torn Berbasis Android dan Mikrokontroller," *JITSI J. Ilm. Teknol. Sist. Inf.*, vol. 1, no. 3, pp. 97–107, 2020.

[14] Tukadi, W. Widodo, M. Ruswiensari, and A. Qomar, "Monitoring Pemakaian Daya Listrik Secara Realtime Berbasis Internet of Things," *Semin. Nas. Sains dan Teknol. Terap. VII 2019*, pp. 581–586, 2018.

[15] M. Faisal, H. Harmadi, and D. Puryanti, "Perancangan Sistem Monitoring Tingkat Kekeruhan Air Secara Realtime Menggunakan Sensor TSD-10," *J. Ilmu Fis. | Univ. Andalas*, vol. 8, no. 1, pp. 9–16, 2016.

[16] M. Kautsar, R. R. Isnanto, and E. D. Widianto, "Sistem Monitoring Digital Penggunaan dan Kualitas Kekeruhan Air PDAM Berbasis Mikrokontroler ATMega328 Menggunakan Sensor Aliran Air dan Sensor Fotodiode," *J. Teknol. dan Sist. Komput.*, vol. 3, no. 1, pp. 79–86, 2015.

[17] J. M. Hudin, D. Susilawati, and M. A. Faisal, "Implementasi Model Agile Pada Monitoring Suhu Kolam Ikan Dengan Algoritma Fuzzy Logic Berbasis Internet of Thing (Iot)," *Swabumi*, vol. 6, no. 2, pp. 133–138, 2018.

[18] A. I. Irawan, R. Patmasari, and M. R. Hidayat, "Peningkatan Kinerja Sensor DS18B20 pada Sistem IoT Monitoring Suhu Kolam Ikan," *JTERA (Jurnal Teknol. Rekayasa)*, vol. 5, no. 1, p. 101, 2020.

[19] Y. Rahmanto, A. Rifaini, S. Samsugi, and S. D. Riskiono, "SISTEM MONITORING pH AIR PADA AQUAPONIK MENGGUNAKAN MIKROKONTROLER ARDUINO UNO," *J. Teknol. dan Sist. Tertanam*, vol. 1, no. 1, p. 23, 2020.

[20] D. Y. Tadeus, K. Azazi, and D. Ariwibowo, "Model Sistem Monitoring pH dan Kekeruhan pada Akuarium Air Tawar berbasis Internet of Things," *Metana*, vol. 15, no. 2, pp. 49–56, 2019.

[21] D. A. Wibisono, S. Aminah, and G. Maulana, "Rancang Bangun Sistem Monitoring Kualitas Air Pada Tambak Udang Berbasis Internet of Things," *Perpust. Univ. Sanata Dharma*, no. September, p. viii, 2019.

e-ISSN: 2614-8404
p-ISSN : 2776-3234

Optimization of Extreme Programming Methods in Plastics Waste Management Company Websites

Linda Perdana Wanti [1*)], Oman Somantri[2], Annisa Romadloni[3], Eka Tripustikasari[4]
[1,2,3]Department of Informatics, Politeknik Negeri Cilacap
[4]Department of Informatics, Universitas Amikom Purwokerto
Email: [1]lindaperdana16@gmail.com, [2]oman.mantri@yahoo.com, [3]annisaromadloni@pnc.ac.id,
[4]ekatripustikasari@amikompurwokerto.ac.id

Abstract – Plastic waste needs to be handled properly according to its type to reduce its negative impact on the earth, such as the issue of global warming which is still being widely discussed among the public. Good and correct plastic waste management has a significant long-term impact on the issue of global warming. Using the optimization of the extreme programming (XP) method to develop a plastic waste management system. With the system development method used, namely extreme programming, this system helps the community to be aware of waste and manage waste as well and wisely as possible. Extreme programming flexibility supports all changes that occur during the process of building this plastic waste management system. The output produced in the construction of this system is the management and sale of plastic waste that can be recycled according to its type. With usability testing that has been carried out, this system has been evaluated and shows a result of 88.07%, this value means that the plastic waste management system is well accepted to be used in plastic waste management.

Keywords – Extreme Programming, Information System, Optimization, Plastics Waste, Usability.

I. INTRODUCTION

The increasing plastic consumption must be wisely and inevitably balanced with its management [1]. Movements to prevent global warming are still often campaigned on social media and others to raise awareness of the importance of correct waste management hence not to add to the burden of the earth [2], [3]. Waste management system is built to identify, separate and manage plastic waste according to type [4], [5]. Good waste management starts from the smallest unit, which is house hold. The waste is collected based on its type, which is recyclable and unrecyclable waste [6], [7]. There will be a further process for recyclable waste, whereas the unrecyclable ones are collected and sorted to its type [8]. Some studies can be used as reference for plastic waste management to reduce the impact of pollution caused by it including [9], which is about recycling plastic waste due to the increasing use of plastic. The use of *Triboelectrostatic* technology by utilizing different surface properties of materials is for these materials to be distinguished in electric charge, deflected in an electric field and collected in separate places. In conclusion, this research recycles plastic therefore it can be reused with *Triboelectrostatic* technology. Other studies by [10] are the utilization of plastic waste with catalytic pyrolysis techniques. The catalytic pyrolysis process itself is a thermochemical decomposition process of organic matter through the process of heating, either using little oxygen or not using oxygen or other chemicals at all. The point is that plastic waste is processed through the pyrolysis process to get new materials that can be utilized. The material can be oil or something else. However some other research examples use plastic waste, namely [11], [12], [13] and [14].

The process of developing a plastic waste management information system use extreme programming method. Optimization of extreme programming methods is to accommodate all changes from the identification and planning process, the system design process, the process of implementing the design into the system, including coding, the system testing process using the testing method and implementing the system to the user [15]. Extreme programming approach is widely used because this method is able to anticipate some problems for example when the system is too fast in making changes, whereas the needs are not clearly defined [16]. The use of extreme programming methods in several studies that have been carried out include [17] implements the XP method in web applications for job training participant selection. The XP method is used to anticipate web development which consists of a small team. Furthermore, in [18], the XP method is used in making customer service complaints applications with university as the research objects. The XP method is able to accommodate all the needs and application development processes relatively quickly with minimal team members. Research by [19], applied the XP method to an online sales system where the system was built with a relatively fast deadline [20]. The impact of using XP method approach to the system which is being built is that the customer does not take a long time during online reservation [21]. However some other research examples use plastic waste, namely [22],

The difference between this research and previous research is that this research optimizes the extreme programming method approach in the construction of a plastic waste management system is used so that the system can be utilized well, considering the system was built with a small team and underwent several considerable changes in a fast time during the building process [23]. Some of the advantages of the XP method are the most optimized and effective, according to the state of the plastic waste management system [24]. With this system, the company is helped because it can maximize what the company has to be optimized and from a social point of view it helps reduce environmental impacts due to waste that is not managed properly [10]. The company gets two profits, namely profits from sales and waste management as well as participating in maintaining the balance of the earth from wastes that take a long time to process until the waste can be reused.

II. Research Methodology

The system development method used for managing plastic waste is by optimizing the extreme programming method. The XP method approach in this case creates a plastic waste management system by minimizing the iteration that will later be carried out. Optimization in terms of iteration in the system development process by making the system from scratch has been endeavored to meet all user needs. A small or minimal developer teams is not a significant obstacle because the team's performance is maximized according to their respective parts. Figure 1 explains the detailed process of the extreme programming method.

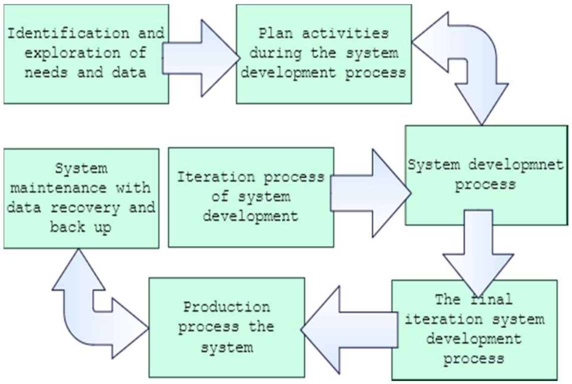

Identification and exploration of needs and data
Plan activities during the system development process
System developmnet process
Iteration process of system development
System maintenance with data recovery and back up
The final iteration system development process
Production process the system

Figure. 1 Extreme Programming Method

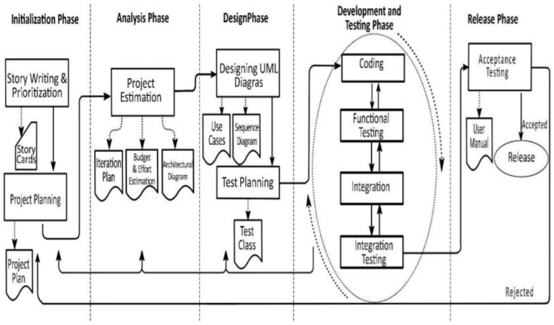


Figure. 2 XP Research Framework

Figure 2 describes the working steps of the extreme programming method which consists of 5 stages/phases. The first step is to initiate end-user needs which are translated and documented into story cards and project planning regarding the activities carried out in the next stages [25]. The second step is the analysis phase. After the system and user requirements are initiated, the results will be analyzed according to the estimated target system development required. The third stage is the design phase, where this phase translates the results of the system requirements analysis into designs such as UML designs, namely class diagrams, activity diagrams, sequence diagrams, use case diagrams and many others [26]. The next stage is to develop the system that was developed and test it with end-users to find out things that are still not in accordance with the needs of end-users and improve them according to feedback and end-users. At this stage, it is necessary to know that it is possible to return to the previous stage in accordance with the direction of improvement from the user. Can goes back to the initiation stage or to the analysis stage or to the design stage. For some projects that have been developed, this iteration usually goes back to the design stage. End users will provide a lot of input and system developers must improve in the system design section. At this stage usability testing is also carried out on the developed system. Usability testing itself is carried out to find out how easy it is for users to use the developed system so that system developers know the end users' difficulties in using the system and find out the shortcomings of the developed system [27]. To determine user satisfaction with the developed system, the measurement uses the following equation:

$$S(\%) = \frac{\sum_{i-1}^{n} X_i}{5*n} * 100\% \qquad (1)$$

Description:
S : Satisfaction
X : Respondent success score
n : Number of respondents

In addition to end-user satisfaction, the level of effectiveness and efficiency of the developed system also needs to be measured to determine the success rate of the system using the user's success rate, which is the percentage of tasks completed correctly by the end-user [28]. To measure the level of system effectiveness and system efficiency using the following equation:

$$Ef\&Es(\%) = \frac{\sum_{i-1}^{n} X_i}{n} * 100\% \qquad (2)$$

Description:
Ef & Es : Efficiency & Effectiveness
X : Respondent success score
n : Number of respondents

Meanwhile, to measure the usability of the system to determine the average of the effectiveness, efficiency and satisfaction of end-users using the equation below [29] :

$$U(\%) = U\ (\%) = \frac{Ef+Es+S}{3} * 100\% \qquad (3)$$

Description:
U : Usability
Ef : Efficiency
Es : Effectiveness
S : Satisfaction

The last stage is the release phase, where at this stage, the system has reached the stage of being reproduced and ready to be implemented in many places according to the initial plan [30]. The system has reached the final stage and there is no further improvement from the end-user.

III. Results And Discussion

It starts with defining and exploring all processes and needs as well as the data that will be used in system development [31]. The second stage is planning the activities which are carried out and the data that has been collected. They are then defined and grouped according to their needs. Next is the system development process with a minimum iteration process which is conditioned and planned at the beginning of the system development process which is suited in accordance with user requirements. After the iteration process is carried out and finished, it is followed by the production stage of the plastic waste management system by defining all the designs and coding stages until the system is successfully built and tested by the user. The final step in optimizing the extreme programming method is the process of maintaining the system by both performing system recovery and backing up the system periodically.

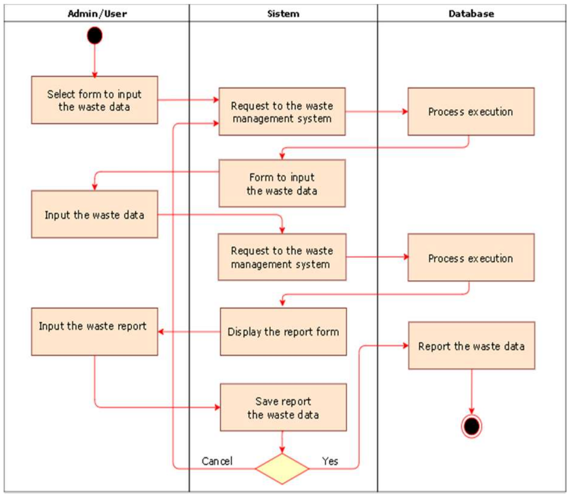

Admin/User
Sistem
Database
Select form to input the waste data
Request to the waste management system
Process execution
Form to input the waste data
Input the waste data
Request to the waste management system
Process execution
Input the waste report
Display the report form
Report the waste data
Save report the waste data
Cancel
Yes

Figure. 3 Activity Diagram of a Plastic Waste Management System

Figure 3 explains the activity diagram of a plastic waste management system. Activity diagram starts from the request form or module user request to the system, where the user requests a form to enter the data of incoming or outgoing waste. The request is responded by the system and the incoming / outgoing waste form is prepared by the database to be transferred to the system and displayed to the user. The user fills in the incoming / outgoing waste form and saves it into the database, and then the request for storing the incoming / outgoing waste form is responded by the system [32]. The system proceeds to the database to execute the stored process for the incoming / outgoing waste form that has been filled. Then the user request to the system to fill in the incoming / outgoing waste report form. The response is responded by the system and forwarded to the database. The database prepares the report form for incoming / outgoing waste, and after the form is filled in by the user, the user saves the form to the database. The storage request is forwarded to the system with the response of the incoming / outgoing waste form has been stored in the system database.

Figure. 4 Main page of the Plastic Waste Management System

Figure 4 displays the main page of the plastic waste management system. There are 4 modules, namely a master for managing data from incoming and outgoing waste, a transaction module for the process of selling and supplying plastic waste and a report module for displaying reports from data input and transaction processes that have been running. Finally, there is exit menu to exit the application.

Usability testing was carried out on the plastic waste management website at PT. HR involved 27 respondents. The respondents consisted of directors, admins and employees of PT. HR, in addition to partners from PT. HR is also a respondent. Usability testing is used to measure respondents' satisfaction with the developed system, namely a plastic waste management website, the effectiveness and efficiency of a website-based system for plastic waste management [33]. For the measurement of self-satisfaction using equation (1) [34]. While the measurement of the level of effectiveness and efficiency of the system uses equation (2) [27]. Previously, the questionnaires distributed to 27 respondents were analyzed, as shown in the following table:

Table 1. Evaluation Elements for the Level of Satisfaction, Effectiveness and Efficiency

Satisfaction Evaluation Element	Effectiveness and efficiency Evaluation Elements
This app is interesting	Scenario to unlock system
This app is easy to use	Scenario for website menu functions
You will suggest friends use this app	Scenario for button function on each form
Reading the text on the screen is very easy	Scenario for master menu
The color composition corresponds to	scenario for partner data input
The image displayed is attractive	scenario for plastic waste data input
The buttons are easy to understand	Scenario for partner transaction menu function
The buttons are easy to use	Scenario for the plastic waste transaction menu function
Website materials are easy to understand	Scenario for report menu function
The language used is easy to understand	Scenario to exit the system

The results of measuring the level of satisfaction, the level of effectiveness and efficiency of the plastic waste management website at PT. HR is done twice. First, before there is input from the end-user, second after the system is repaired on the basis of user feedback for 3 iterations. The first iteration focuses on system design, the second iteration improves the function of the buttons on the menu on the website, and the third iteration improves the system login function. The graph of the results of usability testing is shown in Figure 5 below:

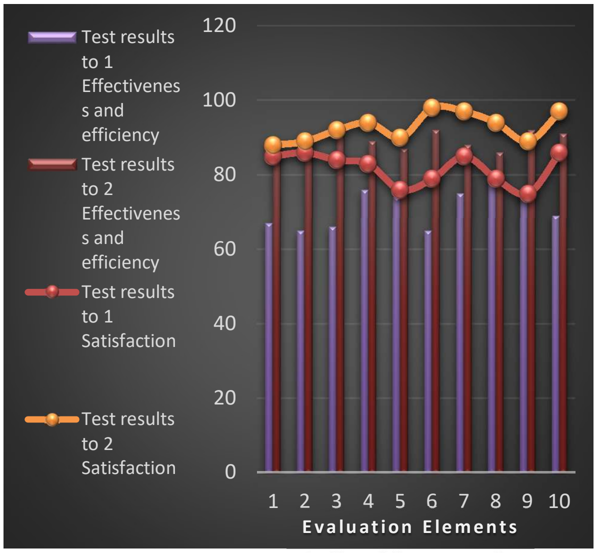


Figure. 5 Main page of the Plastic Waste Management System

The results of the usability of the website-based plastic waste management system based on the average satisfaction level of the 1st and 2nd tests as well as testing the 1st and 2nd level of effectiveness and efficiency of the system using equation (3) are as follows:

$$U\ (\%) = \frac{89.2+86.7+88.3}{3} * 100\% = 88.07\%$$

IV. Conclusion

From the research that has been done, the output produced is a plastic waste management website. Utilization of this website is also used to minimize the impact of plastic waste which is increasingly uncontrollable. A plastic waste management website was built and developed by implementing and optimizing the extreme programming method that is easily adapted for the construction of a system with a relatively short time and significant changes [10]. Iterations were carried out during the construction of the plastic waste management website 3 times by accommodating all feedback from end-users to improve the plastic waste management website according to end-user needs. Usability testing is 88.07%, this shows that the website-based plastic waste management system can be accepted by users to be used and can be used by companies to manage incoming and outgoing plastic waste.

References

[1] Zhang, H., Pap, S., Taggart, M. A., Boyd, K. G., James, N. A., & Gibb, S. W. (2019). A review of the potential utilisation of plastic waste as adsorbent for removal of hazardous priority contaminants from aqueous environments. *Environmental Pollution*, *xxxx*, 113698. https://doi.org/10.1016/j.envpol.2019.113698

[2] Li, X., Ling, T. C., & Hung Mo, K. (2020). Functions and impacts of plastic/rubber wastes as eco-friendly aggregate in concrete – A review. *Construction and Building Materials*, *240*, 117869. https://doi.org/10.1016/j.conbuildmat.2019.117869

[3] Anuar Sharuddin, S. D., Abnisa, F., Wan Daud, W. M. A., & Aroua, M. K. (2016). A review on pyrolysis of plastic wastes. *Energy Conversion and Management*, *115*, 308–326. https://doi.org/10.1016/j.enconman.2016.02.037

[4] Kaliyavaradhan, S. K., Ling, T. C., Guo, M. Z., & Mo, K. H. (2019). Waste resources recycling in controlled low-strength material (CLSM): A critical review on plastic properties. *Journal of Environmental Management*, *241*(December 2018), 383–396. https://doi.org/10.1016/j.jenvman.2019.03.017

[5] Li, W. C., Tse, H. F., & Fok, L. (2016). Plastic waste in the marine environment: A review of sources, occurrence and effects. *Science of the Total Environment*, *566–567*, 333–349. https://doi.org/10.1016/j.scitotenv.2016.05.084

[6] Mwanza, B. G., & Mbohwa, C. (2017). Drivers to Sustainable Plastic Solid Waste Recycling: A Review. *Procedia Manufacturing*, *8*(October 2016), 649–656. https://doi.org/10.1016/j.promfg.2017.02.083

[7] Wedayani, N. M. (2018). Studi Pengelolaan Sampah Plastik Di Pantai Kuta Sebagai Bahan Bakar Minyak. *Jurnal Presipitasi : Media Komunikasi Dan Pengembangan Teknik Lingkungan*, *15*(2), 122. https://doi.org/10.14710/presipitasi.v15i2.122-126

[8] Moharir, R. V., & Kumar, S. (2019). Challenges associated with plastic waste disposal and allied microbial routes for its effective degradation: A comprehensive review. *Journal of Cleaner Production*, *208*, 65–76. https://doi.org/10.1016/j.jclepro.2018.10.059

[9] Wu, G., Li, J., & Xu, Z. (2013). Triboelectrostatic separation for granular plastic waste recycling: A review. *Waste Management*, *33*(3), 585–597. https://doi.org/10.1016/j.wasman.2012.10.014

[10] Miandad, R., Barakat, M. A., Aburiazaiza, A. S., Rehan, M., & Nizami, A. S. (2016). Catalytic pyrolysis of plastic waste: A review. *Process Safety and Environmental Protection*, *102*, 822–838. https://doi.org/10.1016/j.psep.2016.06.022

[11] Al-Harahsheh, M., Al-Nu'Airat, J., Al-Otoom, A., Al-Hammouri, I., Al-Jabali, H., Al-Zoubi, M., & Abu Al'Asal, S. (2019). Treatments of electric arc furnace dust and halogenated plastic wastes: A review. *Journal of Environmental Chemical Engineering*, *7*(1), 102856. https://doi.org/10.1016/j.jece.2018.102856

[12] A., G. K., K., A., M., H., K., S., & G., D. (2019). Review on plastic wastes in marine environment – Biodegradation and biotechnological solutions. *Marine Pollution Bulletin*, *May*, 110733. https://doi.org/10.1016/J.MARPOLBUL.2019.110733

[13] Saleem, J., Adil Riaz, M., & Gordon, M. (2018). Oil sorbents from plastic wastes and polymers: A review. *Journal of Hazardous Materials*, *341*, 424–437. https://doi.org/10.1016/j.jhazmat.2017.07.072

[14] Horodytska, O., Valdés, F. J., & Fullana, A. (2018). Plastic flexible films waste management – A state of art review. *Waste Management*, *77*, 413–425. https://doi.org/10.1016/j.wasman.2018.04.023

[15] Tolfo, C., & Wazlawick, R. S. (2008). The influence of organizational culture on the adoption of extreme programming. *Journal of Systems and Software*, *81*(11), 1955–1967. https://doi.org/10.1016/j.jss.2008.01.014

[16] Schneider, J. G., & Johnston, L. (2005). eXtreme Programming - Helpful or harmful in educating undergraduates? *Journal of Systems and Software*, *74*(2 SPEC. ISS.), 121–132. https://doi.org/10.1016/j.jss.2003.09.025

[17] Supriyatna, A. (2018). Metode Extreme Programming Pada Pembangunan Web Aplikasi Seleksi Peserta Pelatihan Kerja. *Jurnal Teknik Informatika*, *11*(1), 1–18. https://doi.org/10.15408/jti.v11i1.6628

[18] Azdy, R. A., & Rini, A. (2018). Penerapan Extreme Programming dalam Membangun Aplikasi Pengaduan Layanan Pelanggan (PaLaPa) pada Perguruan Tinggi. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(2), 197. https://doi.org/10.25126/jtiik.201852658

[19] Gumelar, T., Astuti, R., & Sunarni, A. T. (2017). Sistem Penjualan Online Dengan Metode Extreme Programming. *Jurnal Telematika*, *9*(2), 87–90.

[20] Anwer, F., & Aftab, S. (2017). SXP: Simplified Extreme Programing Process Model. *International Journal of Modern Education and Computer Science*, *9*(6), 25–31. https://doi.org/10.5815/ijmecs.2017.06.04

[21] Wanti, L. P., Ikhtiagung, G. N., & Pangestu, I. A. (2021). *Implementasi Extreme programming Pada Website Marketplace Lapak Petani Online*. *12*(01), 50–58. https://doi.org/10.35970/infotekmesin.v12i1.427

[22] Fojtik, R. (2011). Extreme programming in development of specific software. *Procedia Computer Science*, *3*, 1464–1468. https://doi.org/10.1016/j.procs.2011.01.032

[23] Rahmi, R., Sari, R., & Suhatman, R. (2016). Pendekatan Metodologi Extreme Programming pada Aplikasi E-Commerce (Studi Kasus Sistem Informasi Penjualan Alat-alat Telekomunikasi). *Jurnal Komputer Terapan*, *2*(2), 83–92.

[24] Roky, H., & Meriouh, Y. Al. (2015). Evaluation by Users of an Industrial Information System (XPPS) Based on the DeLone and McLean Model for IS Success. *Procedia Economics and Finance*, *26*(0), 903–913. https://doi.org/10.1016/s2212-5671(15)00903-x

[25] Tabassum, A., Manzoor, I., Shahid, D., Rida, A., & Imtiaz, D. (2017). Optimized Quality Model for Agile Development: Extreme Programming (XP) as a Case Scenario. *International Journal of Advanced Computer Science and Applications*, *8*(4), 392–400. https://doi.org/10.14569/ijacsa.2017.080453

[26] Anwer, F., Aftab, S., Shah, S., & Waheed, U. (2017). Comparative analysis of two popular agile process models: extreme programming and scrum. *International Journal of Computer Science and Telecommunications*, *8*(2), 1–7.

[27] Doerfler, R. M., Diamantidis, C. J., Wagner, L. A., Scism, B. M., Vaughn-Cooke, M., Fink, W. J., Blakeman, T., & Fink, J. C. (2019). Usability testing of a sick-day protocol in CKD. *Clinical Journal of the American Society of Nephrology*, *14*(4), 583–585. https://doi.org/10.2215/CJN.13221118

[28] Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & McTear, M. (2019). Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? *ECCE 2019 - Proceedings of the 31st European Conference on Cognitive Ergonomics: "'Design for Cognition,'"* 207–214. https://doi.org/10.1145/3335082.3335094

[29] Korableva, O., Durand, T., Kalimullina, O., & Stepanova, I. (2019). Usability testing of MOOC: Identifying user interface problems. *ICEIS 2019 - Proceedings of the 21st International Conference on Enterprise Information Systems*, *2*(Iceis), 468–475. https://doi.org/10.5220/0007800004680475

[30] Aldave, A., Vara, J. M., Granada, D., & Marcos, E. (2019). Leveraging creativity in requirements elicitation within agile software development: A systematic literature review. *Journal of Systems and Software*, *157*. https://doi.org/10.1016/j.jss.2019.110396

[31] Wanti, L. P., & Romadhon, S. (2020). Implementasi Forward Chaining Method Pada Sistem Pakar Untuk Deteksi Dini Penyakit Ikan. *Infotekmesin*. https://doi.org/10.35970/infotekmesin.v11i2.248

[32] Suryanto, T. (2018). Penerapan E-Marketplace pada Distro Silver Squad. *Konferensi Nasional Sistem Informasi (KNSI) 2018*, *0*(0), 8–9.

[33] Liu, Y. C., Chen, C. H., Lee, C. W., Lin, Y. S., Chen, H. Y., Yeh, J. Y., & Chiu, S. Y. H. (2016). Design and usability evaluation of user-centered and visual-based aids for dietary food measurement on mobile devices in a randomized controlled trial. *Journal of Biomedical Informatics*, *64*, 122–130. https://doi.org/10.1016/j.jbi.2016.10.001

[34] Russ, A. L., Jahn, M. A., Patel, H., Porter, B. W., Nguyen, K. A., Zillich, A. J., Linsky, A., & Simon, S. R. (2018). Usability evaluation of a medication reconciliation tool: Embedding safety probes to assess users' detection of medication discrepancies. *Journal of Biomedical Informatics*, *82*, 178–186. https://doi.org/10.1016/j.jbi.2018.05.002

DDoS ATTACK MITIGATION WITH INTRUSION DETECTION SYSTEM (IDS) USING TELEGRAM BOTS

Mohammad Taufan Asri Zaen[1*)], Ahmad Tantoni[2], Maulana Ashari[3]
[1,3]Program Studi Studi Sistem Informasi, STMIK Lombok
[2]Program Studi Teknik Informatika, STMIK Lombok
Email: [1]opanzain@gmail.com, [2]ahmad.tantoni@students.amikom.ac.id, [3]aarydarkmaul@gmail.com

Abstract – In the current IS/IT era, service to consumers is an absolute must to be prepared to survive in business competition. Physical and logical attacks with the aim of disrupting information technology services for individuals/agencies/companies or reducing the performance of IS/IT used. The development of IoT in the industrial revolution 4.0, which is all online, is a challenge in itself, from a negative point of view, all of them are able to carry out attacks on ISP servers, often carried out by hackers. DDoS (Distributed Denial of Service) attacks are the most common attacks. The development of software for DDoS attacks is very much on the internet, including UDP Unicorn software to attack very easily and can be done by anyone. Software for real-time monitoring of DDoS attacks, one of which is the Telegram bot. Telegram is a messaging system centered on security and confidentiality, while bots are computer programs that do certain jobs automatically. Telegram bot is free, lightweight and multiplatform. In the case study, this research contains 10 access points to the internet that will be mitigated from DDoS attacks. In this study, it was found that DDoS attacks caused traffic to become very high/congested by fulfilling upload traffic so that legitimate traffic users could not access the internet, connection to the internet was slow, the traffic was also unnatural, making it unable to connect to wireless devices and making Mikrotik loginpage becomes unable to appear. The purpose of this study is to mitigate DDoS attacks with the help of telegram bots so as to facilitate the notification of DDoS attacks in the event of an attack so that it is fast to deal with and find the perpetrators of the attack. The conclusion of this study is that DDoS attacks using UDP unicorn software resulted in a traffic spike of 53.5 Mbps on the upload traffic side, causing traffic for legitimate/authenticated users to slow down. By using telegram bots to know DDoS attacks occur in real time with a success rate of attack detection up to 100% notifications on telegram bots. Mitigation of DDoS attacks takes steps to track users using the torch feature on the routerboard interface menu, trace internet connection lines using wired or wireless transmission media, and ensure always monitoring the proxy interface from winbox.

Keywords – Attack Mitigation, DDoS, IDS, Telegram

I. INTRODUCTION

The need for network security is very important in the world of information technology and information systems. In the current IS/IT era, service to consumers is an absolute must to be prepared to survive in business competition. There are times when irresponsible people make attacks on information technology systems and networks that are developed. The attack is in the form of physical and logical attacks with the aim of disrupting information technology services for individuals/agencies/companies or reducing the performance of information systems and information technology used.

The development of IoT (Internet of Things) in the industrial revolution 4.0 which is all online is a challenge in itself, everyone can access anything from the virtual world, from the negative side, everyone is able to attack a service on the internet. Several attacks on servers as internet service providers (ISPs) are often carried out by hackers with various purposes. DDOS (Distributed Denial of Service) attack is an attack that may often be found among other attacks.

The development of software to carry out DDoS attacks is also very widely spread on the internet, including UDP Unicorn software, which is a software that uses a very easy way to attack and can be done by anyone, and there are many more software to carry out DDoS attacks on the internet.

In monitoring computer networks when a DDoS attack occurs, there are many applications that can be used to monitor DDoS attacks in real time, one of which is the telegram bot. Telegram is a cross-platform messaging system centered on security and privacy, while bots are computer programs that do certain jobs automatically.

Telegram bot is a bot that is currently popularly used among the public because it is free, lightweight and multiplatform. Telegram also has a fairly complete and growing Bot API. The famous telegram bot is the telegram-bot made by Yago Perez [1].

In the case study, this research contains 10 access points to the internet that will be mitigated from DDoS attacks. In this study, it was found that DDoS attacks caused traffic to become very high/congested by fulfilling upload traffic so that legitimate traffic users could not access the internet, connection to the internet was slow, the traffic was also unnatural, making it unable to connect to wireless devices and making Mikrotik loginpage becomes unable to appear.

The problem of this research how to mitigate DDoS attacks carried out by using a bot Unicorn UDP telegram. The purpose of this study is to mitigate DDoS attacks with the help of telegram bots so as to facilitate the notification of DDoS attacks in the event of an attack so that it is fast to deal with and find the perpetrators of the attack.

Research conducted by Nadila Sugianti et al with the title Detection of HTTP-Based Distributed Denial of Services (DDOS) Attacks Using the Fuzzy Sugeno Method. The purpose of this research is to create an application to detect HTTP-based DDOS attacks with good accuracy using the fuzzy Sugeno method. Based on the discussion that has been explained and the results of tests that have been carried out, HTTP-based DDOS attack detectors based on the number of users, the number of packets, the number of packets/users and the length of the data captured by fuzzy logic using the Sugeno method can be used as a detector in determining DDOS attacks based on HTTP with an accuracy rate of up to 90% [2].

Research conducted by Jodi Chris Jordan Sihombing et al with the title Implementation of Distributed Denial of Service (DDoS) Attack Detection and Mitigation System using SVM Classifier on Software Defined Network (SDN) Architecture. The research objective is to implement a system that can detect and mitigate DDoS attacks on the SDN architecture. Conclusions from the research 1) The DDoS attack detection and mitigation (SDMD) system using the SVM classifier can be applied to the SDN architecture. 2) The DDoS attack mitigation mechanism is carried out by adding a flow rule on the switch to filter packets that go to the victim host. After the flow rule is added to the switch's flow table, the switch will drop every packet originating from the attacker's source IP, but any packets originating from the legitimate host's source IP will be forwarded. 3) SDMD performance in detecting DDoS attacks is very good. The accuracy obtained in detecting DDoS attacks is 96.08%, 95.66%, and 98.76% for syn flooding, udp flooding, and icmp flooding, respectively [3].

Research conducted by Muhammad Aziz et al with the title Implementing Artificial Neural Networks to Detect DDoS Attacks in Network Forensics. The purpose of the study is to determine the accuracy of DDoS attacks for network forensic purposes, the proposed method to analyze and test DDoS attacks detected on IDS with datasets at the Research Laboratory of Masters in Informatics Engineering, Ahmad Dahlan University (LRis-MTIUAD) using artificial neural network (ANN) methods based on calculations statistics. The conclusion from the research is that the attack information that has been detected by signature-based IDS needs to be reviewed for accuracy using classification with statistical calculations. Based on the analysis and testing carried out by the artificial neural network method, it was found that the accuracy was 95.2381%. Artificial neural network methods can be applied in the field of network forensics in determining accurate results and helping strengthen evidence at trial [4].

Research conducted by Eddy Prasetyo Nugroho et al with the title Security Reporting System on Cloud Computing Networks Through Telegram bots using Intrusion Detection and Prevention System Techniques. The conclusion of the research is to build an intrusion detection system by producing output not only in the form of records of intrusion activities on the database but also notifications via instant messaging Telegram. The results of recording intrusion activities in the database are used as data to perform network forensic analysis regarding the identity of the source of incoming packets, as well as to analyze the level of IDS system responsibility both in detecting attacks and sending notifications to administrators [5].

The research conducted by Jefree Fahana et al with the title Using Telegram as an Attack Notification for Network Forensics Purposes. The purpose of the research is to help network administrators to make it easier to find attacks that are usually carried out manually. The conclusion of the study was that it was successful in detecting attacks by using Snort. Alerts work very well and are able to send information to the database which is then forwarded using the telegram instant messenger application in real time. The results show that there has been a ddos attack via ICMP based on the log analysis performed [6].

1.1 Intrusion Detection System (IDS)

Intrusion Detection System is a prevention system using software or hardware that works automatically to monitor the situation on a computer network and can analyze network security problems. IDS is a tool, method and resource that provides assistance in identifying, reporting on computer network activity [7].

The ability of the IDS is to provide an early warning to the Network Administrator when a certain activity occurs that the Administrator does not want. In addition to providing warnings, IDS is also able to track activities that harm a system. An IDS can monitor packets passing through the network and try to find out if there is any suspicious activity.

IDS functions to monitor unusual activities on the network so that the initial steps of the attackers can be known. Thus the Administrator can take precautions and be prepared for what might happen.

In recognizing attack patterns, there are several methods of how IDS works, namely: Signature Based IDS and Anomaly Based IDS.

a) Signature Based IDS

A signature-based IDS will monitor the packets in the network and compare these packets with the signature database owned by this IDS system. This method is almost the same as how antivirus applications work in detecting malware. The point is that there will be a delay between the detection of an attack on the internet and the signature used for detection which is implemented in the IDS database used. So it could be that the signature database used in the IDS system is not able to detect an attempted attack on the network because the information on this type of attack is not contained in the signature database of this IDS system. During this time delay, the IDS system cannot detect any new types of attacks.

b) Anomaly Based IDS

This type will monitor the traffic in the network and compare the traffic that occurs with the average traffic (stable). The system will identify what is meant by a "normal" network in the network, how much bandwidth is usually used on the network, what protocols are used, what ports are usually interconnected with each other in the network and alert the administrator. when detected something is not normal.

The anomaly-based IDS method offers advantages over the signature-based IDS, which is that it can detect new forms of attacks that are not yet included in the IDS signature database. The downside is that this type often emits false positive messages. So that the Administrator's task becomes more complicated, by having to sort out which is the real attack from the many false positives that appear.

1.2 Distributed Denial of Service

According to Mousavi (2014), Distributed Denial of Service attack (DDoS) is an attack carried out on all the bloat or computer network by sending the dense traffic [8]. DDoS attacks start from attackers who distribute attacks using machines. During an attack, all traffic is directed to the victim's computer or server to consume the victim's resources. DDoS attacks will often use IP spoofing with the aim of flooding the target with high traffic while masking the identity of the original source to prevent mitigation efforts. If the source IP address is spoofed and continues to be scrambled, attempts to block attacks will become difficult. According to Kumarasamy (2012), there are several types of DDoS attacks [9] including the following :

1. TCP SYN Flooding: Is a type of attack that exploits the 3 way handshake mechanism in the TCP protocol, the attacker sends a large number of SYN packets to overload the target being attacked.
2. UDP Flooding: Is a type of attack that exploits the UDP protocol, the attacker has a list of broadcast addresses to send fake UDP packets to. This packet delivery mechanism is sent to a random port and then changes the target location unexpectedly.
3. Ping (ICMP) Flooding: Is a type of attack against the ICMP protocol with the aim of depleting the victim's computer resources by flooding it through requests from ICMP echo or also known as the ping command.

1.3 Firewall Mikrotik

A firewall is a hardware device or software system or group of systems (routers, proxies, gateways) designed to allow or deny network transmissions based on a set of security rules and regulations to enforce control between two networks to protect the "inside" network from the "outside" network. Firewall acts as a filter between internal and external computers. The firewall performs control based on the source IP address, source and destination TCP/UDP ports, destination IP address, and header information stored in data packets [10].

1.4 Traffic Monitoring Mikrotik

Traffic Monitor is a feature in Mikrotik that is rarely used. This traffic monitor can be used to monitor traffic running on an interface on the router and can determine a traffic threshold value. If the traffic has reached the specified threshold, then Traffic Monitor can execute a script. Thus can use this feature for various needs by determining what scripts will be executed. Steps to activate Traffic Monitor in the required interface with the steps on the Tools menu → Traffic Monitor [11].

1.5 Telegram

The main attraction of Telegram is that it can run on various devices and operating systems, not only mobile phones, but also computers, smartphones and others. Telegrams and bots can make everyday life easier without having to just stare at the computer. At the beginning of the development of the bot world on Telegram, almost all bots were created using telegram-cli and lua. The telegram-cli bot works like a personal account, it can even log in as a telegram-cli bot account and do what normal accounts can do. The benefits of this bot were also acknowledged by Telegram, which then launched a bot API so that many people could build bots using the programming language they mastered without having to deal with Telegram-cli or MTProto. Bot API is a bot account, there are certain things that normal accounts can do that bot accounts can't, for example creating groups, adding people to groups and removing people from groups [12].

II. RESEARCH METHODOLOGY

2.1 Research Flow

The flow of this research is as follows :

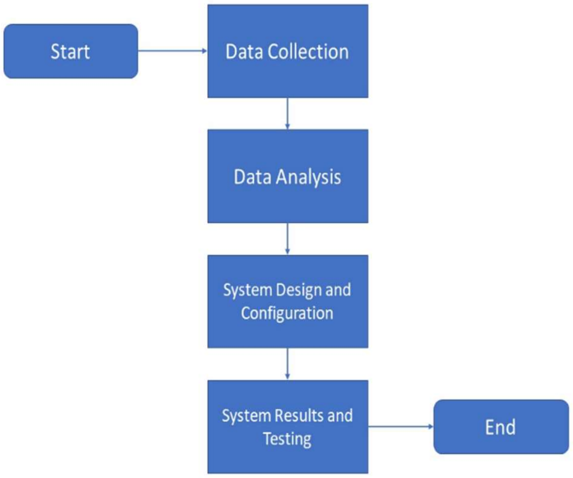


Figure 1. Research Flow

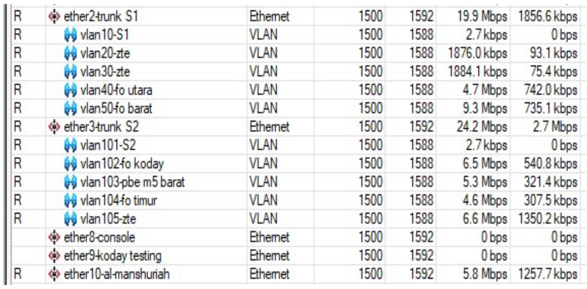
Figure 1. Shows the flow of the research carried out. The first step is data collection, which collects data on the form and DDoS attacks that are carried out. The second step is data analysis which performs analysis and techniques to prevent DDoS attacks. the third step is the design and configuration of the system which designs and implements DDoS attack mitigation techniques with telegram bots as notifications in the event of an ongoing attack. The fourth step is the results and system testing which is doing testing to find out whether the DDoS attack mitigation system runs smoothly as expected.

2.2 Data collection

Data collection with the tested parameters is to determine the target IP address, determine the packet size to be sent (flood), the path that is passed (threads), the socket path that is passed (sockets per thread) and the test time for 10 seconds ago. UDP unicorn software is run, then monitoring is carried out from the mikrotik routerboard interface.

2.3 Data analysis

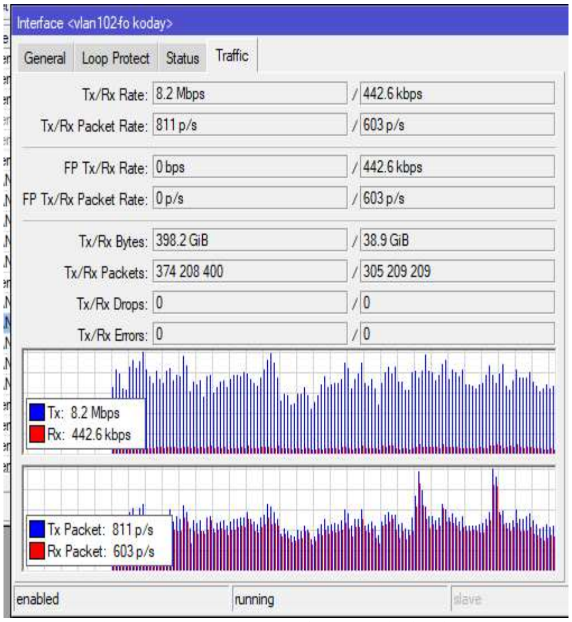
Data analysis at the location of DDoS attacks carried out for research is as follows :



R	ether2-trunk S1	Ethernet	1500	1592	19.9 Mbps	1856.6 kbps
R	vlan10-S1	VLAN	1500	1588	2.7 kbps	0 bps
R	vlan20-zte	VLAN	1500	1588	1876.0 kbps	93.1 kbps
R	vlan30-zte	VLAN	1500	1588	1884.1 kbps	75.4 kbps
R	vlan40-fo utara	VLAN	1500	1588	4.7 Mbps	742.0 kbps
R	vlan50-fo barat	VLAN	1500	1588	9.3 Mbps	735.1 kbps
R	ether3-trunk S2	Ethernet	1500	1592	24.2 Mbps	2.7 Mbps
R	vlan101-S2	VLAN	1500	1588	2.7 kbps	0 bps
R	vlan102-fo koday	VLAN	1500	1588	6.5 Mbps	540.8 kbps
R	vlan103-pbe m5 barat	VLAN	1500	1588	5.3 Mbps	321.4 kbps
R	vlan104-fo timur	VLAN	1500	1588	4.6 Mbps	307.5 kbps
R	vlan105-zte	VLAN	1500	1588	6.6 Mbps	1350.2 kbps
	ether8-console	Ethernet	1500	1592	0 bps	0 bps
	ether9-koday testing	Ethernet	1500	1592	0 bps	0 bps
R	ether10-al-manshuriah	Ethernet	1500	1592	5.8 Mbps	1257.7 kbps

Figure 2. Mikrotik interface

Figure 2. Shows 10 active mikrotik interfaces or 10 internet access sharing lines including 1).vlan20-ZTE, 2).vlan30-ZTE, 3).vlan40-north fo, 4).vlan50-west fo, 5).vlan102- fo koday, 6).vlan103-pbe m5 west, 7).vlan104-fo east, 8).vlan105-zte, 9).ether9-koday testing and lastly 10).ether10-al-manshuriah. These 10 paths are the paths that come out of mikrotik.



Tx/Rx Rate: 8.2 Mbps / 442.6 kbps
Tx/Rx Packet Rate: 811 p/s / 603 p/s
FP Tx/Rx Rate: 0 bps / 442.6 kbps
FP Tx/Rx Packet Rate: 0 p/s / 603 p/s
Tx/Rx Bytes: 398.2 GiB / 38.9 GiB
Tx/Rx Packets: 374 208 400 / 305 209 209
Tx/Rx Drops: 0 / 0
Tx/Rx Errors: 0 / 0

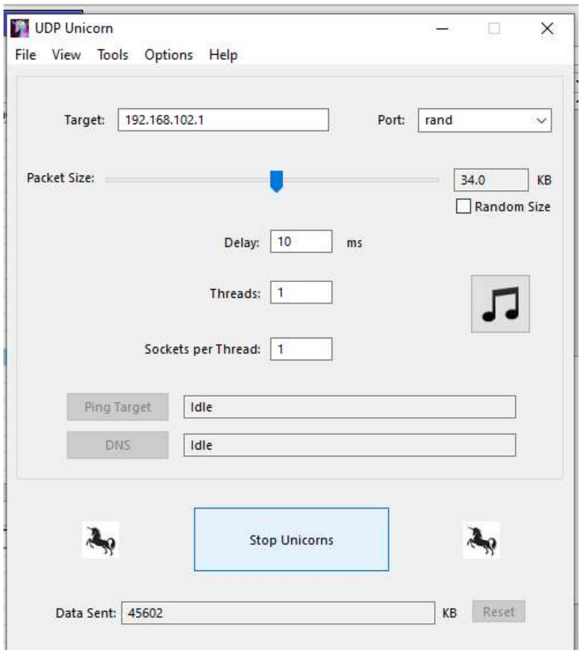
Figure 3. Interface <vlan102-fo koday>

Figure 3. Shows the <vlan102-fo koday> interface showing normal traffic and no DDoS attacks have occurred. The average download traffic is 8.2Mbps and the average upload traffic is 44.6kbps. The download result is greater than the upload result.

2.4 Unicorn UDP Attack Form

Figure 4. Shows UDP unicorn software by entering the server's IP address in the target column and providing the value of the packet size to be sent then clicking start unicorn to run a DDoS attack on the target

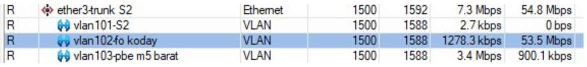
The forms of UDP Unicorn software attacks when used to carry out DDoS attacks are as follows :



Target: 192.168.102.1
Port: rand
Packet Size: 34.0 KB
□ Random Size
Delay: 10 ms
Threads: 1
Sockets per Thread: 1
Ping Target: Idle
DNS: Idle

Data Sent: 45602 KB

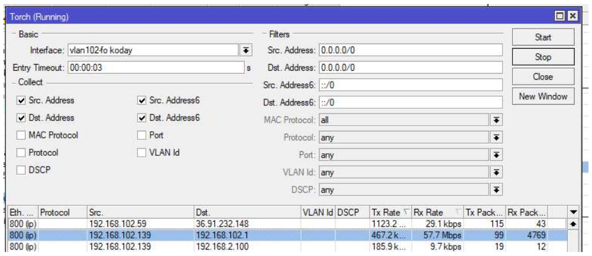
Figure 4. UDP Unicorn



R	ether3-trunk S2	Ethernet	1500	1592	7.3 Mbps	54.8 Mbps
R	vlan101-S2	VLAN	1500	1588	2.7 kbps	0 bps
R	vlan102-fo koday	VLAN	1500	1588	1278.3 kbps	53.5 Mbps
R	vlan103-pbe m5 barat	VLAN	1500	1588	3.4 Mbps	900.1 kbps

Figure 5. Unicorn UDP Attack

Figure 5. Shows an attack from the UDP unicorn software by filling up the upload traffic so that legitimate user traffic cannot access the internet. When an attack occurs, the monitored download traffic is 1278.3kbps while the monitored upload traffic is 53.5Mbps. An increase in unnatural upload traffic on this interface causes the internet connection to be slow, unable to connect to wireless devices, loginpage cannot appear.



Basic
Interface: vlan102-fo koday
Entry Timeout: 00:00:03 s
Collect
☑ Src. Address
☑ Dst. Address
□ MAC Protocol
□ Protocol
□ DSCP
☑ Src. Address6
☑ Dst. Address6
□ Port
□ VLAN Id
Filters
Src. Address: 0.0.0.0/0
Dst. Address: 0.0.0.0/0
Src. Address6: ::/0
Dst. Address6: ::/0
MAC Protocol: all
Protocol: any
Port: any
VLAN Id: any
DSCP: any

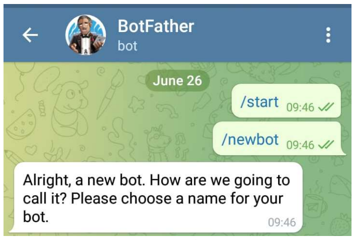
Eth. ...	Protocol	Src.	Dst.	VLAN Id	DSCP	Tx Rate	Rx Rate	Tx Pack...	Rx Pack...
800 (ip)		192.168.102.59	36.91.232.148			1123.2 ...	29.1 kbps	115	43
800 (ip)		192.168.102.139	192.168.102.1			467.2 k...	57.7 Mbps	99	4769
800 (ip)		192.168.102.139	192.168.2.100			185.9 k...	9.7 kbps	19	12

Figure 6. Mikrotik Torch Menu

Figure 6. Shows the torch menu on the Mikrotik router with the aim of knowing the IP address of the DDoS attacker by clicking start then shorting the highest Rx Rate. Figure 6 gets very high upload traffic with an average upload of 57.7Mbps.

2.5 DDoS Attack Detection and Mitigation System Design

The design of the DDoS attack detection and mitigation system using the telegram bot features as follows :



BotFather
bot

June 26

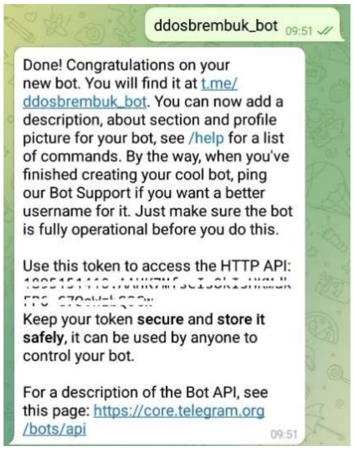
/start 09:46

/newbot 09:46

Alright, a new bot. How are we going to call it? Please choose a name for your bot. 09:46

Figure 7. BotFather Telegram

Figure 7. Shows the BotFather telegram that will be used as a telegram bot machine by writing the commands "/start" and "/newbot" to create a bot script.



ddosbrembuk_bot 09:51

Done! Congratulations on your new bot. You will find it at t.me/ddosbrembuk_bot. You can now add a description, about section and profile picture for your bot, see /help for a list of commands. By the way, when you've finished creating your cool bot, ping our Bot Support if you want a better username for it. Just make sure the bot is fully operational before you do this.

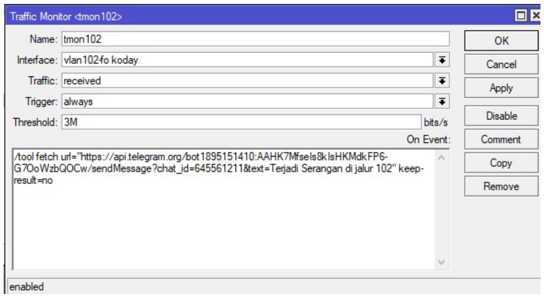
Use this token to access the HTTP API:
[illegible]

Keep your token **secure** and **store it safely**, it can be used by anyone to control your bot.

For a description of the Bot API, see this page: https://core.telegram.org/bots/api 09:51

Figure 8. BotFather Telegram

Figure 8. Shows confirmation of the name of the bot, namely "ddosbrembuk_bot" which will be created to get a token to access the HTTP API telegram.



Traffic Monitor <tmon102>

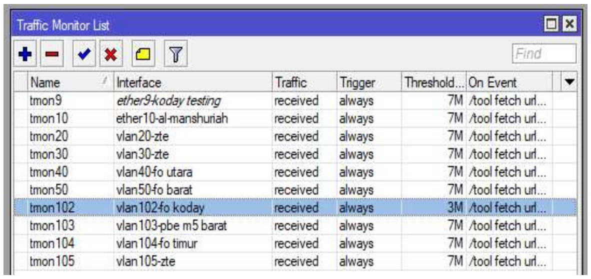
Name: tmon102
Interface: vlan102-fo koday
Traffic: received
Trigger: always
Threshold: 3M bits/s
On Event:

```
/tool fetch url="https://api.telegram.org/bot1895151410:AAHK7Mfsels8kIsHKMdkFP6-
G7OoWzbQOCw/sendMessage?chat_id=645561211&text=Terjadi Serangan di jalur 102" keep-
result=no
```

enabled

Fifure 9. Trafiic Monitor <tmon102>

Figure 9. Shows the existing monitor traffic on the proxy feature. With this traffic monitor it helps manual work when monitoring a DDoS attack that makes traffic very high. This monitor traffic menu is found in the Mikrotik tools when opening Winbox. The name column fills in the name tmon102 in order to provide a different name. The interface column selects the interface to be monitored. The traffic column uses received with the aim of monitoring traffic only on uploads. The trigger column uses always with the aim of always reporting all monitoring traffic activities. The threshold column provides a value of 3Mbps with the aim that if the upload traffic exceeds 3Mbps, it will report to the Telegram bot. The on event column adds a telegram bot script by entering the telegram bot token that has been obtained in the previous stage. The script will display a notification in real time on telegram with the message "An attack occurred on Line 102"



Traffic Monitor List

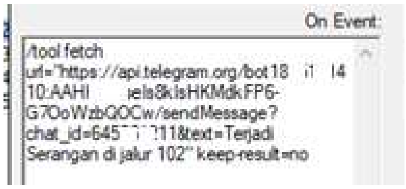
Name	Interface	Traffic	Trigger	Threshold...	On Event
tmon9	*ether9-koday testing*	received	always	7M	/tool fetch url...
tmon10	ether10-al-manshuriah	received	always	7M	/tool fetch url...
tmon20	vlan20-zte	received	always	7M	/tool fetch url...
tmon30	vlan30-zte	received	always	7M	/tool fetch url...
tmon40	vlan40-fo utara	received	always	7M	/tool fetch url...
tmon50	vlan50-fo barat	received	always	7M	/tool fetch url...
tmon102	vlan102-fo koday	received	always	3M	/tool fetch url...
tmon103	vlan103-pbe m5 barat	received	always	7M	/tool fetch url...
tmon104	vlan104-fo timur	received	always	7M	/tool fetch url...
tmon105	vlan105-zte	received	always	7M	/tool fetch url...

Figure 10. Traffic Monitor List

Figure 10. Shows a list of monitor traffic. In the picture there are 10 internet access sharing lines that will be monitored for upload traffic to mitigate ongoing DDoS attacks. The results of this script will immediately make a real-time notification to Telegram in the event of a DDoS attack. In the name column, fill in the name and adjust it to the monitored path so that it provides a different name and in the threshold column it adjusts to the maximum desired upload speed.

III. RESULTS AND DISCUSSION

The results and discussion of DDoS attack mitigation using telegram are as follows :

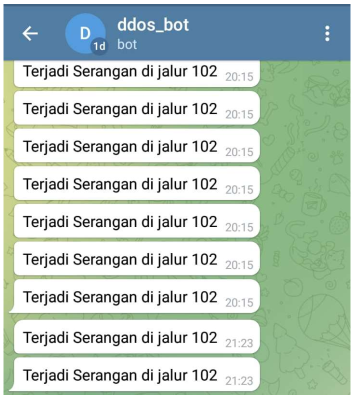


On Event:

```
/tool fetch
url="https://api.telegram.org/bot18 i1 14
10:AAHI jels8kIsHKMdkFP6-
G7OoWzbQOCw/sendMessage?
chat_id=645 11&text=Terjadi
Serangan di jalur 102" keep-result=no
```

Figure 11. Script On Event

Figure 11. shows this script generated with BotFather from the telegram bot by forwarding the url and chat_id links embedded in the script, as follows: url=https://api.telegram.org/bot18951410:AAHK7MfseIs8kIsP6-G7OoWzbw, chat_id=645211. Then give a comment or notification in the form of text = An attack occurred on line 102.



ddos_bot
bot

Terjadi Serangan di jalur 102 20:15

Terjadi Serangan di jalur 102 20:15

Terjadi Serangan di jalur 102 20:15

Terjadi Serangan di jalur 102 20:15

Terjadi Serangan di jalur 102 20:15

Terjadi Serangan di jalur 102 20:15

Terjadi Serangan di jalur 102 20:15

Terjadi Serangan di jalur 102 21:23

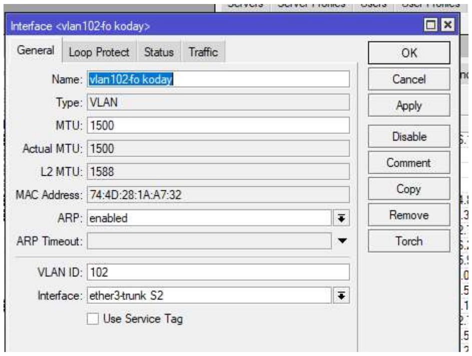
Terjadi Serangan di jalur 102 21:23

Figure 12. DDoS Attack Notification Results

Figure 12. Shows the results of realtime DDoS attack notifications when there is a spike in upload data traffic. If there is data traffic that exceeds the value above 3Mbps, a notification will be sent to the Telegram application according to the 3Mbps upload/threshold value setting in Figure 9. If there are notifications that

enter the Telegram application more than 10 times within one second, it means that a DDoS attack has occurred.

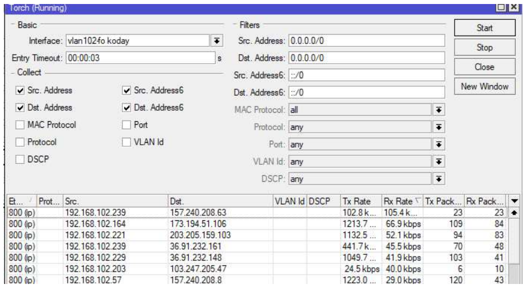
After knowing which internet connection line is being attacked, the first step is to immediately take action by tracing the user using the torch feature on the Mikrotik routerboard interface menu.



Name: vlan102fo koday
Type: VLAN
MTU: 1500
Actual MTU: 1500
L2 MTU: 1588
MAC Address: 74:4D:28:1A:A7:32
ARP: enabled
ARP Timeout:
VLAN ID: 102
Interface: ether3-trunk S2
□ Use Service Tag

Figure 13. Unlock Torch Features

Figure 13. Shows how to open the torch feature by clicking the interface menu → then clicking the torch feature.



Interface: vlan102fo koday
Entry Timeout: 00:00:03 s
☑ Src. Address
☑ Src. Address6
☑ Dst. Address
☑ Dst. Address6
□ MAC Protocol
□ Port
□ Protocol
□ VLAN Id
□ DSCP
Src. Address: 0.0.0.0/0
Dst. Address: 0.0.0.0/0
Src. Address6: ::/0
Dst. Address6: ::/0
MAC Protocol: all
Protocol: any
Port: any
VLAN Id: any
DSCP: any

Et...	Prot...	Src.	Dst.	VLAN Id	DSCP	Tx Rate	Rx Rate	Tx Pack...	Rx Pack...
800 (ip)		192.168.102.239	157.240.208.63			102.8 k...	105.4 k...	23	23
800 (ip)		192.168.102.164	173.194.51.106			1213.7 ...	66.9 kbps	109	84
800 (ip)		192.168.102.221	203.205.159.103			1132.5 ...	52.1 kbps	94	83
800 (ip)		192.168.102.239	36.91.232.161			441.7 k...	45.5 kbps	70	48
800 (ip)		192.168.102.229	36.91.232.148			1049.7 ...	41.9 kbps	103	41
800 (ip)		192.168.102.203	103.247.205.47			24.5 kbps	40.0 kbps	6	10
800 (ip)		192.168.102.57	157.240.208.8			1223.0 ...	29.0 kbps	120	43

Figure 14. Torch Feature

Figure 14. Shows the torch feature, with this feature it is easy to see data traffic (Tx/Rx) which can be sorted from the largest to the smallest. To see the attacker by looking at the source address and sorting the upload data traffic (Rx) from the largest to the smallest.

Then with the second step, trace the internet connection path using wired or wireless transmission media by removing the RJ45 connector connected to each computer network device, and ensuring always monitoring the Mikrotik interface from Winbox.

IV. CONCLUSION

The conclusion of this research is: 1) There was a DDoS attack using UDP unicorn software which resulted in a traffic spike of 53.5 Mbps on the upload traffic side which resulted in slow traffic of legitimate/authenticated users. 2) By using telegram bots to know DDoS attacks occur in real time with a success rate of attack detection up to 100% notifications on telegram bots. Mitigation of DDoS attack mitigation as soon as possible take steps a) take action to track users using the torch feature on the Mikrotik routerboard interface menu by clicking the interface menu where the attack occurred and then clicking the Torch feature. b) trace the internet connection path using wired or wireless transmission media by removing the rj45 connector connected to each computer network device (router/switch), as well as ensuring always monitoring the proxy interface from winbox.

REFERENCES

[1] Kabayankababayan, “Mengenal Bot Telegram,” 2015. https://rizaumami.github.io/2015/12/11/mengenal-bot-telegram/ (accessed Dec. 01, 2021).

[2] N. Sugianti, Y. Galuh, S. Fatia, and K. F. H. Holle, “Deteksi Serangan Distributed Denial of Services (DDOS) Berbasis HTTP Menggunakan Metode Fuzzy Sugeno,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 4, no. 3, pp. 156–164, 2020, doi: 10.14421/jiska.2020.43-03.

[3] J. C. J. Sihombing, D. P. Kartikasari, and A. Bhawiyuga, “Implementasi Sistem Deteksi dan Mitigasi Serangan Distributed Denial of Service (DDoS) menggunakan SVM Classifier pada Arsitektur Software-Defined Network (SDN),” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 10, pp. 9608–9613, 2019.

[4] M. Aziz, R. Umar, and F. Ridho, “Implemetasi Jaringan Saraf Tiruan untuk Mendeteksi Serangan DDoS pada Forensik Jaringan,” *QUERY J. Sist. Inf.*, vol. 3, no. 1, pp. 46–52, 2019.

[5] E. P. Nugroho, E. Nugraha, and M. N. Zulfikar, “Sistem Reporting Keamanan pada Jaringan Cloud Computing Melalui bot Telegram dengan Menggunakan Teknik Intrussion Detection and Prevention System,” *J. Teknol. Terpadu*, vol. 5, no. 2, pp. 49–57, 2019, [Online]. Available: https://journal.nurulfikri.ac.id/index.php/JTT/article/view/233.

[6] J. Fahana, R. Umar, and F. Ridho, “Pemanfaatan Telegram sebagai Notifikasi Serangan untuk Jaringan Forensik,” *Query J. Inf. Syst.*, vol. 1, no. 2, pp. 6–14, 2017, [Online]. Available: http://jurnal.uinsu.ac.id/index.php/query/article/view/1036.

[7] D. Ariyus, *INTRUSION DETECTION SYSTEM: Sistem Deteksi Penyusupan Pada Jaringan Komputer*. Yogyakarta: Andi, 2007.

[8] S. M. Mousavi, “Early Detection of DDoS Attacks in Software Defined Networks Controller,” in *Thesis*, Ottawa, 2014, pp. 77–81.

[9] S. kumarasamy and R. Asokan, “Distributed Denial of Service (DDOS) Attacks Detection Mechanism,” *Int. J. Comput. Sci. Eng. Inf. Technol.*, vol. 1, no. 5, pp. 39–49, 2011, doi: 10.5121/ijcseit.2011.1504.

[10] A. Chopra, “Security Issues of Firewall,” *Int. J. P2P Netw. Trends Technol.*, vol. 22, no. 1, pp. 4–9, 2016, doi: 10.14445/22492615/ijptt-v22p402.

[11] citraweb, “Traffic Monitor Mikrotik,” *mikrotik.id*. https://mikrotik.id/artikel_lihat.php?id=289 (accessed Aug. 11, 2021).

[12] S. R. Umami, “Mengenal Bot Telegram,” *2015*, 2015. https://rizaumami.github.io/2015/12/11/mengenal-bot-telegram/ (accessed Aug. 11, 2021).

e-ISSN: 2614-8404
p-ISSN:2776-3234

Application of Information Gain to Select Attributes in Improving Naive Bayes Accuracy in Predicting Customer's Payment Capability

Herfandi[1*], Mohammad Taufan Asri Zaen[2], Yuliadi[3], M. Julkarnain[4], Fahri Hamdani [5]
[1,3,4,5] Teknik Informatika, Universitas Teknologi Sumbawa, Sumbawa, Indonesia
[2] Sistem Informasi STMK Lombok, Lombok Tengah, Indonesia
Email: [1]herfandi@uts.ac.id, [2]opanzain@gmail.com, [3,]yuliadi@uts.ac.id, [4]m.julkarnain@uts.ac.id, [5]fahri.hamdani@uts.ac.id

Abstract – The customer is the main factor in the running of PT. XYZ. A good understanding of customers is very important for predicting the capability of customers to pay. The implementation of credit collectibility is used to determine the quality of customer credit, one of which is the customer's capability to pay interest and principal on time. While manually, it is very difficult to accurately predict the capability of customer credit payments. Data mining techniques with the Naïve Bayes algorithm were chosen to classify customers to be able to find patterns, analyze and predict, because they have good performance, are efficient, and simple. The Naïve Bayes algorithm has a weakness in terms of sensitivity to many attributes, so the accuracy is low. Based on the problem stated, his study will apply the Information Gain method to select the most influential attribute on the label in order to increase the accuracy of the Naïve Bayes algorithm. This research produces a new dataset with seven attributes: TENOR, SALARY, DOWN PAYMENT, INSTALLMENT, APPROVAL, OTR CLASS, AGE with Labels: Status and Id: Id number based on the Information Gain method. The dataset comparison process with 995 data records showed an increase in accuracy, precision, and AUC using the new dataset compared to the old dataset, but in the t-Test test with an alpha value = 0.05 there is a difference but not significant. In the evaluation process, performance experienced a significant increase in the use of new datasets with the following percentages of performance improvement: accuracy = 8%, precision = 18.42%, recall = 17.65% and AUC= 0.057%. The results of this study obtained AUC of 0.876, accuracy of 87.88%, precision of 61.90%, and recall of 76.47%, and classified into good classification.

Keywords: credit collectibility, customer prediction, Data Mining, Naive Bayes, Information Gain

I. INTRODUCTION

In the present era, business strategy has developed very significantly. Customers occupy a very important position in this regard. Customers are also a major factor in the running of PT. XYZ. A good understanding of customers is very important for predicting the capability of customers to pay in the future. The implementation of credit collectibility is used to determine the quality of customer credit[1], one of which is the customer's capability to pay interest and principal on time that has been mutually agreed upon[2]. The problem currently being faced is that manually it is very difficult to predict the capability of customer credit payments accurately based on the dataset they have[3], and many companies have difficulty identifying customers who are able to pay on time[4].

The data mining technique using the customer classification approach is an approach that is widely used to find patterns, analyze and predict[5]. With a customer classification approach based on existing datasets, we can predict a customer's payment capability[6]. Meanwhile, manually, it is very difficult to accurately predict the capability of customer credit payments based on the dataset they have.

Currently, there are many data mining techniques with classification approach algorithms that have been used to find patterns, analyze and predict customer behavior, such as the Decision Tree Algorithm[7], Neural Network[8], Support Vector Machine[9], Naive Bayes[10], and K-Nearest Neighbor [11]. The Naïve Bayes algorithm is the algorithm with the most widely used classification approach and was chosen to classify customers because it has good performance, is efficient, and is simple in terms of finding patterns, analyzing and predicting[12]. The Naïve Bayes algorithm has one drawback, namely that it is sensitive to many attributes, so the accuracy is low. Selection or choosing the attribute that has the most influence on the label is very important for the Nave Bayes algorithm to be able to increase the accuracy of the algorithm[13].

One of the methods for determining the best attribute or selecting the attribute that has the most influence on the label is the Information Gain method. The Information Gain method is superior to other methods because the Information Gain method will measure how much absence and presence of an attribute that plays a role in making good classification decisions in any class or label. The Information Gain method is one of the successful attribute selection approaches in classification[14].

In this study, the authors apply the Information Gain method to select the most influential attribute on the label to be applied to the new dataset in an effort to improve the accuracy of the Naïve Bayes Algorithm in predicting the payment capability of customers at PT. XYZ.

1.1 Collectibility of Credit

Credit collectibility is a credit quality status that is taken from a person's score or track record in the banking world[15]. This quality is based on 3 main standards, one of which is the customer's capability to pay principal and interest on time that has been mutually agreed upon[16]. Low collectibility or bad loans can affect the economic condition of a business and worsen the trickle down effect on the overall economy, where this has an impact on the company's growth and income in the future[17]. There are several important elements in the provision of a credit facility, namely Trust, Agreement, Term, Risk and Rewards[18]. To find out how to provide credit to customers based on good credit quality, it is necessary to accurately predict the capability of customer credit payments as a reference for management in making decisions to improve credit quality and collectability[19].

1.2 Data Mining

Data mining is a process of using existing or past data, then processing it so that it finds patterns, meaningful relationships, and trends by examining a set of stored data using statistical and mathematical techniques[20]. Data mining became popular in the 1990s as a solution for extracting previously unknown patterns and information based on a set of data[21]. Data mining can complete several jobs and is divided into four groups, namely prediction modeling, cluster analysis, association analysis, and anomaly detection[22]. However, data mining techniques can also be applied to other data representations, such as spatial, text-based, and multimedia (image) data domain[23]. Data Mining can also be defined as the process of extracting information from large data sets through the use of algorithms with techniques taken from the field of statistics and Database Management Systems[24].

II. RESEARCH METHODOLOGY

The object of this research is the Population History database of customer payments in the New Motorcycle (NMC) program at PT. XYZ year 2019-2020. The use of population data is expected to be able to find out what kind of customer criteria can complete credit on time. Software instrument used are:

1) Delphi 7 : Programming language
2) MySQL : Database system software
3) Rapid Miner 9.9 : Tools that help in testing
4) SQLYog : Database administrator application

2.1 Research Process Flow

The research process flow adopts the CRISP-DM (Cross Standard Industries Process for Data Mining) model, where the CRIPS-DM model is modified according to the research stages[25]:

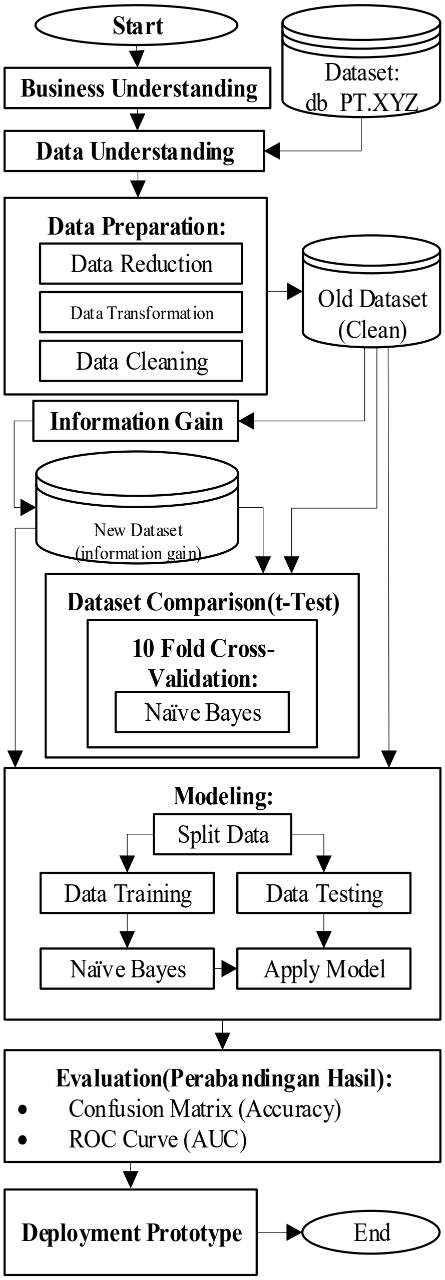


Figure 1. Research Process Flow

The flow of the research process described can be explained as follows:

a. Business Understanding: at this stage, the researcher understands the problem in the research object and then looks for solutions and goals to solve it.

b. Understanding data: at this stage, the researcher determines and collects what data is needed and then defines it according to the solution and research objectives.

c. Data Preparation: at this stage, the researcher cleans the data so that he gets a clean dataset that will be used as a classification model.

d. Information gain: researchers will choose the best or most influential attribute on the label and then make it a new dataset.

e. Dataset Comparison: at this stage, the researcher compares the old dataset and the new dataset on the Naïve Bayes algorithm by means of a different test using t-Test.

f. Modeling: at this stage, the researcher will model the data based on the dataset and a new dataset with a 90% data split (Training) : 10% (Testing)

g. Evaluation: at this stage, the researcher will compare the results of measuring accuracy, precision, recall, and AUC on the old dataset and the new dataset.

h. Development: this stage, the researcher builds a prototype that will be used to predict the customer's payment capability.

2.2 Research Formulas

a. Naïve Bayes Algorithm

The Naive Bayes algorithm is an algorithm with a classification approach for predicting a simple probabilistic based that was put forward by the English scientist Thomas Bayes which is based on the application of Bayes' theorem or rules with the assumption of strong independence on features, meaning that a feature in a data is not related to the existence or the absence of other features in the same data[26].

The advantage of using the Nave Bayes algorithm classification approach is that the Nave Bayes algorithm only requires a small amount of training data needed for the classification process[27]. The Naïve Bayes algorithm classification approach has been proven to be applicable in real and complex situations[27].

The Naïve Bayes algorithm classification approach can be defined as follows[28]:

$$\boldsymbol{P(c|x) = \frac{P(x|c)P(c)}{P(x)}} \tag{1}$$

where

P(c|x) : posterior probability of the label (target) to the predictor (attribute)

P(x|c) : likelihood, which is the probability of the predictor on the label

P(c) : prior probability of the label

P(x) : prior probability of the predictor.

b. The Information Gain Method

The Information Gain method is a method of selecting or selecting attributes in the simplest dataset by only ranking the attributes. The Information Gain method is widely used in the application of data analysis categorization, text, microarray, and image data analysis[29]. In the selection of dataset attributes, the pattern classification approach plays a very important role[30]. The Information Gain method can help reduce noise caused by irrelevant attributes[31]. The initial step that must be done is to determine the best attribute value by calculating the entropy value. Entropy is the process of using the probability of certain events or attributes to measure class uncertainty[32]. After calculating the entropy value, then we can only calculate the Information Gain method[33].

Calculating entropy is defined as follows[33]:

$$\boldsymbol{Entropy(S) = \sum_{i}^{c} - P_i \log_2 P_i} \tag{2}$$

where c is the number of values on the classification label and Pi is the number of samples for class i.

The information gain method is defined as follows [23]:

$$\boldsymbol{Gain\ (S, A) = Entropy(S) - \sum\nolimits_{Values\ (A)} \frac{|S_v|}{S} Entropy(S_v)} \tag{3}$$

where A is an attribute, v is a possible value for attribute A, Values(A) is the set of possible values for A, |Sv| is the number of samples for the value of v, |S| is the sum of all data samples and Entropy (Sv) is the entropy for samples that have a value of v.

2.3 Prototype Development Design

The following is the flow of the Flowchart which is implemented into the prototype payment capability prediction:

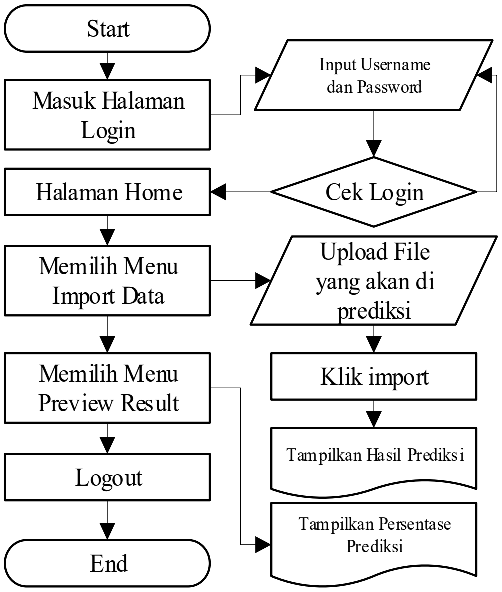

Start
Masuk Halaman Login
Input Username dan Password
Cek Login
Halaman Home
Memilih Menu Import Data
Upload File yang akan di prediksi
Klik import
Memilih Menu Preview Result
Tampilkan Hasil Prediksi
Logout
Tampilkan Persentase Prediksi
End

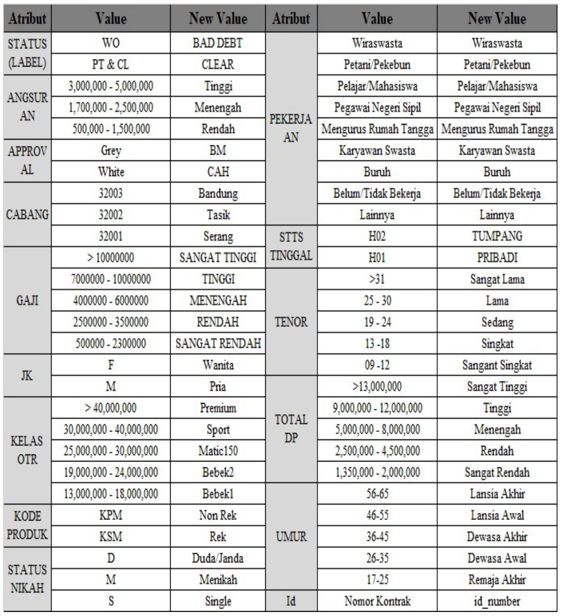
Figure 2. Flowchart Prototype

The user opens the application where the application will directly direct the user to the login page menu, the user enters the username and password to enter the home page, to predict the new data, the user selects the import data menu and uploads the data to be predicted. The user clicks import to see the prediction result display. After that, the user can choose the preview result menu to see the percentage of the predicted results from the data, to exit the user logs out.

III. RESULTS AND DISCUSSION

This research begins with the business understanding stage, where researchers find the problem manually is very difficult to accurately predict the capability of customer credit payments based on the dataset they have, as well as the Naïve Bayes algorithm which is sensitive to many attributes. The next stage is data understanding. At this stage, the researcher combines data from 4 tables, namely active management summary contracts, order management application hiders, order management applications, and acctmgmt.ar contracts. From this process, researchers get 995 data from 2019-2020. The data taken is contract data from 3 branch offices, namely Serang Branch, Bandung Branch, and Tasik Branch, with a total of 27 attributes and 1 label. Furthermore, data preparation, at this stage, the researcher performs data reduction, namely the selection of attributes that are relevant to the target to be achieved. The selected attributes are expected to be determinants of the information to be processed. After data reduction is done, 13 attributes are generated, 1 ID and 1 label. The data transformation stage here is used to obtain a suitable representation for the specific task to be performed. For example: the age value of 17-25 becomes the new value for late teenagers, as can be seen in Table 1.

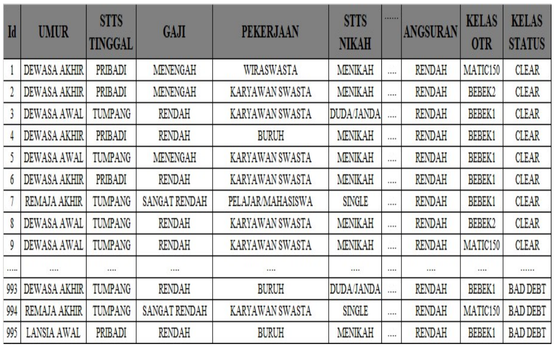
Table 1. Data Transformation



Atribut	Value	New Value	Atribut	Value	New Value
STATUS (LABEL)	WO	BAD DEBT	PEKERJAAN	Wiraswasta	Wiraswasta
	PT & CL	CLEAR		Petani/Pekebun	Petani/Pekebun
ANGSURAN	3,000,000 - 5,000,000	Tinggi		Pelajar/Mahasiswa	Pelajar/Mahasiswa
	1,700,000 - 2,500,000	Menengah		Pegawai Negeri Sipil	Pegawai Negeri Sipil
	500,000 - 1,500,000	Rendah		Mengurus Rumah Tangga	Mengurus Rumah Tangga
APPROVAL	Grey	BM		Karyawan Swasta	Karyawan Swasta
	White	CAH		Buruh	Buruh
CABANG	32003	Bandung		Belum/Tidak Bekerja	Belum/Tidak Bekerja
	32002	Tasik		Lainnya	Lainnya
	32001	Serang	STTS TINGGAL	H02	TUMPANG
GAJI	> 10000000	SANGAT TINGGI		H01	PRIBADI
	7000000 - 10000000	TINGGI	TENOR	>31	Sangat Lama
	4000000 - 6000000	MENENGAH		25 - 30	Lama
	2500000 - 3500000	RENDAH		19 - 24	Sedang
	500000 - 2300000	SANGAT RENDAH		13 -18	Singkat
JK	F	Wanita		09 -12	Sangant Singkat
	M	Pria	TOTAL DP	>13,000,000	Sangat Tinggi
KELAS OTR	> 40,000,000	Premium		9,000,000 - 12,000,000	Tinggi
	30,000,000 - 40,000,000	Sport		5,000,000 - 8,000,000	Menengah
	25,000,000 - 30,000,000	Matic150		2,500,000 - 4,500,000	Rendah
	19,000,000 - 24,000,000	Bebek2		1,350,000 - 2,000,000	Sangat Rendah
	13,000,000 - 18,000,000	Bebek1	UMUR	56-65	Lansia Akhir
KODE PRODUK	KPM	Non Rek		46-55	Lansia Awal
	KSM	Rek		36-45	Dewasa Akhir
STATUS NIKAH	D	Duda/Janda		26-35	Dewasa Awal
	M	Menikah		17-25	Remaja Akhir
	S	Single	Id	Nomor Kontrak	id_number

Then the data cleaning stage is done by filling in the blank data based on the average value, and replacing the data values that do not match. The results of all the stages above produce a dataset (Clean) or old dataset, which can be seen in table 2.

Table 2. Dataset Clean or Old Dataset



Id	UMUR	STTS TINGGAL	GAJI	PEKERJAAN	STTS NIKAH	......	ANGSURAN	KELAS OTR	KELAS STATUS
1	DEWASA AKHIR	PRIBADI	MENENGAH	WIRASWASTA	MENIKAH	....	RENDAH	MATIC150	CLEAR
2	DEWASA AKHIR	PRIBADI	MENENGAH	KARYAWAN SWASTA	MENIKAH	....	RENDAH	BEBEK2	CLEAR
3	DEWASA AWAL	TUMPANG	RENDAH	KARYAWAN SWASTA	DUDA/JANDA	....	RENDAH	BEBEK1	CLEAR
4	DEWASA AKHIR	PRIBADI	RENDAH	BURUH	MENIKAH	....	RENDAH	BEBEK1	CLEAR
5	DEWASA AWAL	TUMPANG	MENENGAH	KARYAWAN SWASTA	MENIKAH	....	RENDAH	BEBEK1	CLEAR
6	DEWASA AKHIR	PRIBADI	RENDAH	KARYAWAN SWASTA	MENIKAH	....	RENDAH	BEBEK1	CLEAR
7	REMAJA AKHIR	TUMPANG	SANGAT RENDAH	PELAJAR/MAHASISWA	SINGLE	....	RENDAH	BEBEK1	CLEAR
8	DEWASA AWAL	TUMPANG	RENDAH	KARYAWAN SWASTA	MENIKAH	....	RENDAH	BEBEK2	CLEAR
9	DEWASA AWAL	TUMPANG	RENDAH	KARYAWAN SWASTA	MENIKAH	....	RENDAH	MATIC150	CLEAR
.....	....	....	....	....	....	....	....	....	......
993	DEWASA AKHIR	TUMPANG	RENDAH	BURUH	DUDA/JANDA	....	RENDAH	BEBEK1	BAD DEBT
994	REMAJA AKHIR	TUMPANG	SANGAT RENDAH	KARYAWAN SWASTA	SINGLE	....	RENDAH	MATIC150	BAD DEBT
995	LANSIA AWAL	PRIBADI	RENDAH	BURUH	MENIKAH	....	RENDAH	BEBEK1	BAD DEBT

3.1 The experimental process

Researchers conducted an experimental process using Rapidminer 9.9 tools to perform Data Cleaning, Information Gain, Dataset Comparison, Modeling, and Evaluation. The selection of Rapidminer tools is considered capable of being used for research, prototyping, and supporting all steps of the data mining process such as data preparation, result visualization, validation, and optimization[34]. The experimental process can be seen in Figure 3.

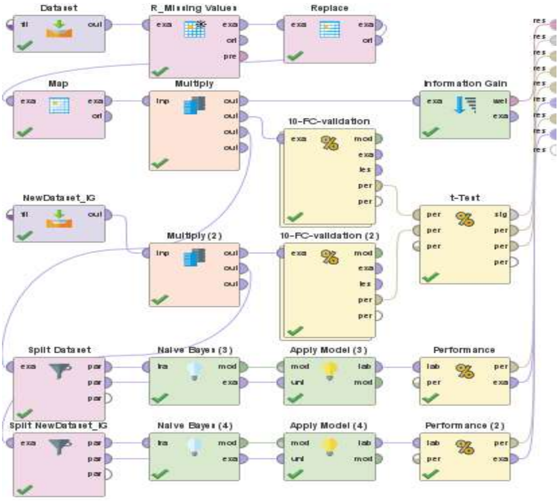


Figure 3. The experiment process

3.2 The Information Gain

The Information Gain method process uses an old dataset of 995 data with specifications of 13 attributes, 1 ID and 1 label to get the best or most influential attribute on the label. The results of the Information Gain Method can be seen in figure 4.

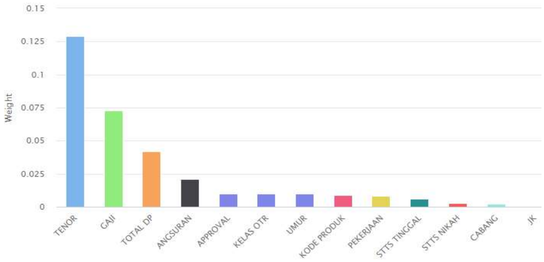


Figure 4. Results of the Information Gain method

The researcher will use the attribute that has the most influence on the label. This attribute will later become an attribute in the new dataset based on the calculation results of the Information Gain method. New attribute by weight method Information Gain can be seen in table 3.

Table 3. New Attributes

No	Attribut	Weight	Status
1	TENOR	0.12898411785188602	new atribut
2	GAJI	0.07251636940706196	new atribut
3	TOTAL DP	0.04205947913917707	new atribut
4	ANGSURAN	0.020693432559648506	new atribut
5	APPROVAL	0.010477564333977618	new atribut
6	KELAS OTR	0.009832172681986773	new atribut
7	UMUR	0.00970525033831704	new atribut
8	KODE PRODUK	0.008562723828346774	delete
9	PEKERJAAN	0.00809223987708052	delete
10	STTS TINGGAL	0.006496154668806264	delete
11	STTS NIKAH	0.0026098786500883264	delete
12	CABANG	0.0020787262163179943	delete
13	JK	0	delete

Based on the calculation results of the Information Gain method in table 3, the new dataset will have the following seven attributes: TENOR, SALARY, DOWN PAYMENT, INSTALLMENT, APPROVAL, OTR CLASS, AGE with Label: Status and Id: Id_number.

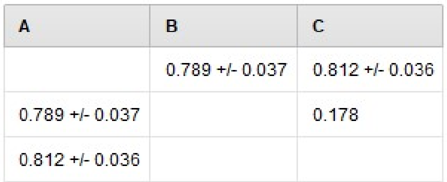
3.3 Dataset Comparison

In this process, the researcher compares the old dataset and the new dataset on the Naïve Bayes algorithm using 10 fold cross-validation by dividing the dataset into 10 parts, of the 10 parts of the data, 9 parts are used as training data, and the remaining 1 part is used as testing data. Then a different test was performed using t-Test to determine the significant difference in the dataset used in the Naïve Bayes algorithm. The results can be seen in the table below.

Table 4. Comparison of 10 fold cross-validation

	Dataset+NB	New_DatasetIG+NB
Accuracy	78.89% +/- 3.70%	81.20% +/- 3.64%
Precision	41.40% +/- 9.38%	45.85% +/- 9.58%
Recall	45.92% +/- 10.39%	44.77% +/- 10.35%
AUC	0.804 +/- 0.066	0.822 +/- 0.055

Table 5. Different Test (t-Test)



A	B	C
	0.789 +/- 0.037	0.812 +/- 0.036
0.789 +/- 0.037		0.178
0.812 +/- 0.036		

Based on the results of the 10-fold cross-validation test, there is an increase in accuracy, precision, and AUC in the Naïve Bayes algorithm using the new dataset compared to the old dataset, but in the t-Test test with an alpha value of 0.05 there is a difference but not significant.

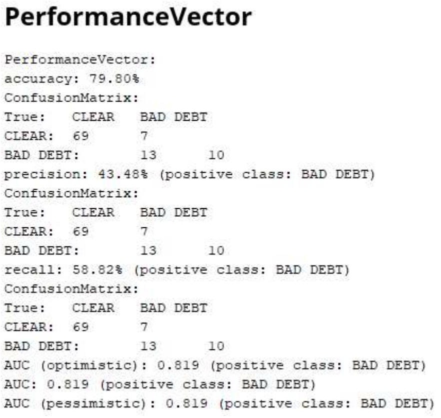
3.4 Modeling

The researcher will model the old dataset and new dataset using the Naïve Bayes algorithm based on the split data operator in rundom subsets with a ratio of 90% (Training): 10% (Testing) as shown in figure 3.

3.5 Evaluation.

This process will compare the model testing of the old dataset and the new dataset that have been determined in the modeling process by measuring their performance. The results of the performance comparison can be seen in table 6.

The following are the results of model testing based on performance measurements and AUC on the old dataset. The results can be seen in figures 5 and 6.



PerformanceVector

```
PerformanceVector:
accuracy: 79.80%
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  69      7
BAD DEBT:       13      10
precision: 43.48% (positive class: BAD DEBT)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  69      7
BAD DEBT:       13      10
recall: 58.82% (positive class: BAD DEBT)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  69      7
BAD DEBT:       13      10
AUC (optimistic): 0.819 (positive class: BAD DEBT)
AUC: 0.819 (positive class: BAD DEBT)
AUC (pessimistic): 0.819 (positive class: BAD DEBT)
```

Figure 5. Performance Vector Old Dataset

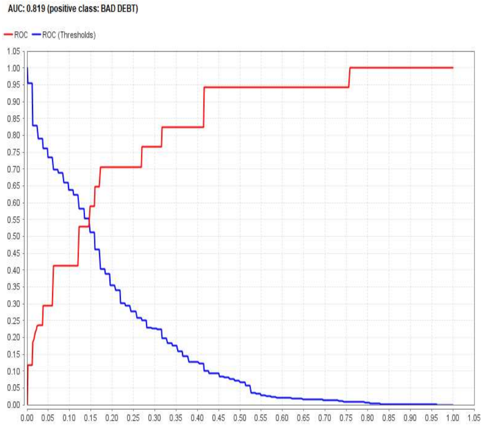
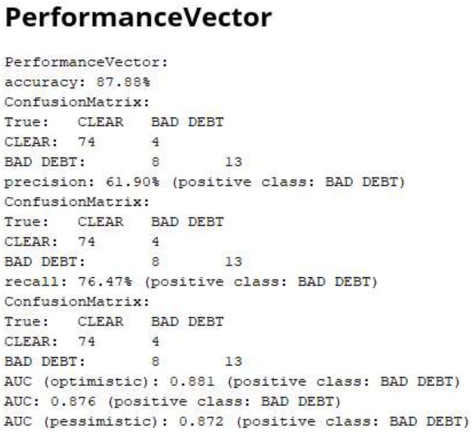


Figure 6. ROC Curve (AUC) Old Dataset

While the results of model testing based on performance measurements and AUC on the new dataset can be seen in figures 7 and 8.



PerformanceVector

```
PerformanceVector:
accuracy: 87.88%
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  74      4
BAD DEBT:       8       13
precision: 61.90% (positive class: BAD DEBT)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  74      4
BAD DEBT:       8       13
recall: 76.47% (positive class: BAD DEBT)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  74      4
BAD DEBT:       8       13
AUC (optimistic): 0.881 (positive class: BAD DEBT)
AUC: 0.876 (positive class: BAD DEBT)
AUC (pessimistic): 0.872 (positive class: BAD DEBT)
```

Figure 7. Performance Vector New Dataset

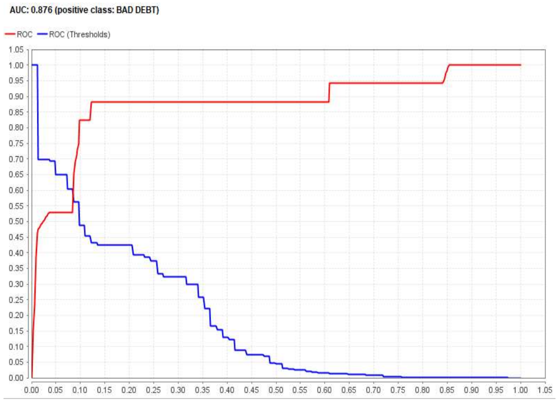


Figure 8. ROC Curve (AUC) New Dataset

Based on the results of the performance and AUC above, the result of the comparison performance is as follows:

Table 6. Result Comparison Performance

Dataset + Algoritma Naïve Bayes			
	true CLEAR	true BAD DEBT	class precision
pred. CLEAR	69	7	90.79%
pred. BAD DEBT	13	10	43.48%
class recall	84.15%	58.82%	
accuracy	79.80%		
precision	43.48% (positive class: BAD DEBT)		
recall	58.82% (positive class: BAD DEBT)		
AUC	0.819 (positive class: BAD DEBT)		
New Dataset (Information Gain) + Algoritma Naïve Bayes			
	true CLEAR	true BAD DEBT	class precision
pred. CLEAR	74	4	94.87%
pred. BAD DEBT	8	13	61.90%
class recall	90.24%	76.47%	
accuracy	87.88%		
precision	61.90% (positive class: BAD DEBT)		
recall	76.47% (positive class: BAD DEBT)		
AUC	0.876 (positive class: BAD DEBT)		

Based on the measurement of the performance model with the operator split data in a rundom subset with a comparison of 90% (Training): 10% (Testing) which can be seen in table 6. The results show that by using a new dataset based on the calculation of the Information Gain method on the Naïve algorithm Bayes is much better than the old dataset, both in terms of Accuracy, Precision, Recall, and AUC. The percentage increase in performance using the new dataset is as follows: Accuracy = 8%, Precision = 18.42%, Recall = 17.65% and AUC = 0.057%.

3.6 Development

This process is based on the results of the dataset comparison and evaluation of the above model testing. It is known that the Naïve Bayes algorithm has a good level of accuracy and performance by using a new dataset based on the calculation of the Information Gain method, so that the rules generated by the Nave Bayes algorithm can be used as rules for making prototypes. The researcher hopes that this prototype can make it easier for PT. XYZ in predicting the capability of customers to pay. for the flowchart prototype can be seen in figure 3.

The prototype used in this study was made desktop-based with programming language using Delphi 7.0 and database using MySQL. The display for the main form of the Graphical User Interface (GUI) prototype predicting the capability of customer credit payments can be seen in the image below.

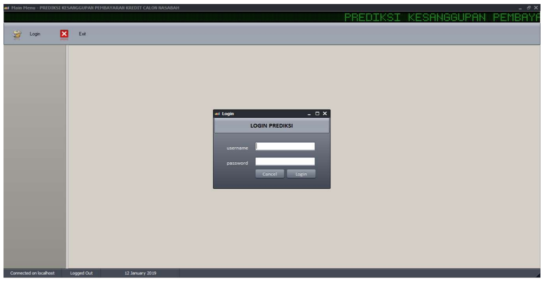
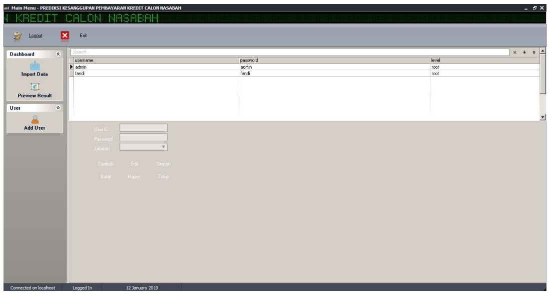


Figure 9. Login Form

When the application is running, the first form will be displayed, which is the login form. For security of access, the user must have a username and password.



username	password	level
admin	admin	root
fandi	fandi	root

Figure 10. Form for Managing Users

The Manage User form is used to view information about registered users, and admins can add, edit, delete users. User accounts can predict new data based on models that have been deployed in the application.

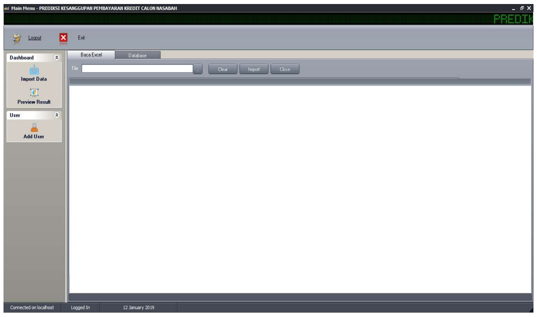


Figure 11. Form File Upload

The import menu is used by the user to make predictions on new data by uploading the data to be predicted, after uploading the user clicks Import to see the display of the prediction results and the user can select the preview result menu to see the percentage number of results from the predicted data.

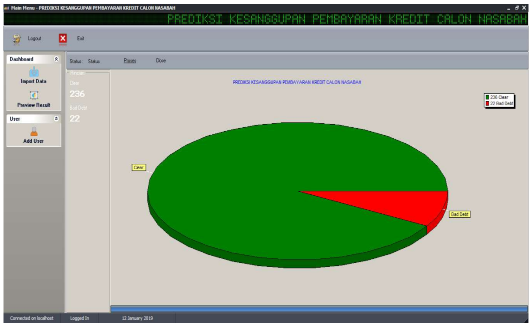


Figure 12. Prediction Result Percentage Form

Testing is carried out with the aim of knowing whether the application is built according to the expected functionality.

Table 7. The Black Box Test

Kelas Uji	Butir Uji	Jenis Pengujian
File Upload	Pilih File	Black Box
	Upload File	Black Box
Dashboard	Lihat Grafik	Black Box

In testing, the upload file is divided into two parts, namely, selecting the file and uploading the file.

Table 8. File Upload Test

Kasus dan Hasil Uji (Data Sesuai)			
File yang di Upload	File dalam bentuk etc.xls	Dapat melakukan pilih file upoload	Diterima
Klick Upload	File berhasil di import oleh sistem	Dapat melakukan upload data	Diterima
Kasus dan Hasil Uji (Data salah)			
Data Masukan	Yang Diharapkan	Pengamatan	Kesimpulan
Extension berbeda, tidak sesuai dengan yang di inginkan sistem	Data Tidak Tersimpan	Data tidak tersimpan dan menampilkan kesalahan	Diterima

In testing the results, the user just clicks on the results.

Table 9. Testing Results

Kasus dan Hasil Uji (Data sesuai)			
Data Masukan	Yang Diharapkan	Pengamatan	Kesimpulan
File yang telah di upload	Berhasil menampilkan Grafik persentase	Dapat menampilkan grafik presentase	Diterima
Klik Tombol Detail	Dapat menampilkan detail customer baik yang disetujui ataupun ditolak	Dapat menampilkan sesuai yang diharapkan	Diterima

Based on the Black Box testing that has been done, it explains that the application that was built has been running well and as expected and The results of this study obtained AUC of 0.876, accuracy of 87.88%, precision of 61.90%, and recall of 76.47%, and classified into good classification.

IV. CONCLUSION

Based on the Research Process Flow that has been carried out by the researchers, it can be concluded:

1. The data preparation process produces 13 attributes, 1 ID and 1 label with a total of 995 data from 2019-2020.
2. The process of calculating the Information Gain method on the old dataset produces a new dataset with seven attributes: TENOR, SALARY, DOWN PAYMENT, INSTALLMENT, APPROVAL, OTR CLASS, AGE and Label: Status and Id: Id_number.
3. In the dataset comparison process, there is an increase in accuracy, precision, and AUC using the new dataset compared to the old dataset, but in the t-Test test with an alpha value of 0.05, there is a difference but not significant.
4. In the evaluation process, performance experienced a significant increase in the use of new datasets with the following percentage increases in performance: Accuracy = 8%, Precision = 18.42%, Recall = 17.65% and AUC = 0.057%.
5. The development process Based on Black Box testing, the application that was built was running well and as expected and The results of this study obtained AUC of 0.876, accuracy of 87.88%, precision of 61.90%, and recall of 76.47%, and classified into good classification.

This study has not been able to provide good t-Test test results because there is no significant difference in the t-Test test results. Based on the findings of this study, it is suggested that further researchers can use chi square, log likelihood ratio or others to select the most influential attribute in order to get good t-test results on the Naïve Bayes algorithm.

REFERENCES

[1] R. Parvizi and M. A. Adibi, “Assessing and Validating Bank Customers Using Data Mining Algorithms for Loan Home,” *Int. J. Ind. Eng. Oper. Res.*, vol. 2, no. 1, 2020.

[2] L. N. Rani, “Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit,” *INOVTEK Polbeng - Seri Inform.*, vol. 1, no. 2, p. 126, 2016, doi: 10.35314/isi.v1i2.131.

[3] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, “Machine learning for email spam filtering: review, approaches and open research problems,” *Heliyon*, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[4] A. Raorane and R. V Kulkarni, “Data Mining Techniques: A Source for Consumer Behavior Analysis,” *Int. J. Database Manag. Syst.*, vol. 3, no. 3, pp. 45–56, 2011, doi: 10.5121/ijdms.2011.3304.

[5] A. U. Khasanah and Harwati, “A Comparative Study to Predict Student’s Performance Using Educational Data Mining Techniques,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 215, no. 1, 2017, doi: 10.1088/1757-899X/215/1/012036.

[6] M. Ala’raj, M. F. Abbod, and M. Majdalawieh, “Modelling customers credit card behaviour using bidirectional LSTM neural networks,” *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00461-7.

[7] M. Sudhakar, C. V. K. Reddy, and A. Pradesh, “TWO STEP CREDIT RISK ASSESMENT MODEL FOR RETAIL BANK LOAN APPLICATIONS USING DECISION TREE DATA MINING TECHNIQUE Research Scholar , D epartment of Computer Science and Technology Professor , D epartment of Physics , Rayalaseema University Kurnool , Andhra,” vol. 5, no. 3, 2016.

[8] P. M. Addo, D. Guegan, and B. Hassani, “Credit risk analysis using machine and deep learning models,” *Risks*, vol. 6, no. 2, pp. 1–20, 2018, doi: 10.3390/risks6020038.

[9] J. Shi and B. Xu, “Credit Scoring by Fuzzy Support Vector Machines with a Novel Membership Function,” *J. Risk Financ. Manag.*, vol. 9, no. 4, p. 13, 2016, doi: 10.3390/jrfm9040013.

[10] A. Krichene, “Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank,” *J. Econ. Financ. Adm. Sci.*, vol. 22, no. 42, pp. 3–24, 2017, doi: 10.1108/JEFAS-02-2017-0039.

[11] Jamaluddin and R. Siringoringo, “Improved Fuzzy K-Nearest Neighbor Using Modified Particle Swarm Optimization,” *J. Phys. Conf. Ser.*, vol. 930, no. 1, 2017, doi: 10.1088/1742-6596/930/1/012024.

[12] F. Harahap, A. Y. N. Harahap, E. Ekadiansyah, R. N. Sari, R. Adawiyah, and C. B. Harahap, “Implementation of Naïve Bayes Classification Method for Predicting Purchase,” *2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018*, no. April, 2019, doi: 10.1109/CITSM.2018.8674324.

[13] H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, and I. Cholissodin, “Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 3, p. 180, 2017, doi: 10.25126/jtiik.201743251.

[14] A. Harris and A. E. Mintaria, “Komparasi

Information Gain , Gain Ratio , CFs-Bestfirst dan CFs-PSO Search Terhadap Performa Deteksi Anomali," vol. 5, pp. 332–343, 2021, doi: 10.30865/mib.v5i1.2258.

[15] D. Dwihandayani, "Analisis Kinerja Non Performing Loan (Npl) Perbankan Di Indonesia Dan Faktor-Faktor Yang Mempengaruhi Npl," *J. Ilm. Ekon. Bisnis*, vol. 22, no. 3, p. 228985, 2017.

[16] U. Al Qoroni, "PROFITABILITAS (Studi pada PT . Federal International Finance Rangkasbitung)," vol. 26, no. 1, pp. 1–5, 2015.

[17] S. Rusnaini, H.- Hamirul, and A. M, "Non Performing Loan (Npl) Dan Return on Asset (Roa) Di Koperasi Nusantara Muara Bungo," *J. Ilm. Manajemen, Ekon. Akunt.*, vol. 3, no. 1, pp. 1–18, 2019, doi: 10.31955/mea.vol3.iss1.pp1-18.

[18] M. Agustiningtyas, "Analisis Faktor-Faktor Yang Mempengaruhi Non Performing Loans Kredit Pada Bank Umum di Indonesia," vol. 1, no. September, pp. 120–133, 2018.

[19] T. T. Muryono and I. Irwansyah, "Implementasi Data Mining Untuk Menentukan Kelayakan Pemberian Kredit Dengan Menggunakan Algoritma K-Nearest Neighbors (K-Nn)," *Infotech J. Technol. Inf.*, vol. 6, no. 1, pp. 43–48, 2020, doi: 10.37365/jti.v6i1.78.

[20] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.

[21] E. Knowledge, "Data Mining : Extracting Knowledge from Data."

[22] J. Sinaga and B. Sinaga, "Data Mining Classification Of Filing Credit Customers Without Collateral With K-Nearest Neighbor Algorithm (Case study: PT. BPR Diori Double)," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 2, no. 2, pp. 204–210, 2020, doi: 10.47709/cnapc.v2i2.401.

[23] M. Otivation and I. Ntroduction, "M Achine L Earning and D Ata M Ining," no. September, pp. 1–21, 2012, doi: 10.13140/RG.2.2.20395.49446/1.

[24] A. J. Chamatkar and P. Butey, "Importance of Data Mining with Different Types of Data Applications and Challenging Areas," *J. Eng. Res. Appl. www.ijera.com*, vol. 4, no. 5, pp. 38–41, 2014.

[25] Konacaklı Enis and KARAARSLAN ENİS, "Artificial Intelligence and Applied Mathematics in Engineering Problems," *Artif. Intell. Appl. Math. Eng. Probl. - Proc. Int. Conf. Artif. Intell. Appl. Math. Eng. (ICAIAME 2019)*, vol. 43, no. January, 2020, doi: 10.1007/978-3-030-36178-5.

[26] S. Karthika and N. Sairam, "A Naïve Bayesian classifier for educational qualification," *Indian J. Sci. Technol.*, vol. 8, no. 16, 2015, doi: 10.17485/ijst/2015/v8i16/62055.

[27] A. P. Wibawa *et al.*, "Naïve Bayes Classifier for Journal Quartile Classification," *Int. J. Recent Contrib. from Eng. Sci. IT*, vol. 7, no. 2, p. 91, 2019, doi: 10.3991/ijes.v7i2.10659.

[28] A. W. Syaputri, E. Irwandi, and M. Mustakim, "Naïve Bayes Algorithm for Classification of Student Major's Specialization," *J. Intell. Comput. Heal. Informatics*, vol. 1, no. 1, p. 17, 2020, doi: 10.26714/jichi.v1i1.5570.

[29] S. Chormunge and S. Jena, "Efficient feature subset selection algorithm for high dimensional data," *Int. J. Electr. Comput. Eng.*, vol. 6, no. 4, pp. 1880–1888, 2016, doi: 10.11591/ijece.v6i4.9800.

[30] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Variable selection for Naïve Bayes classification," *Comput. Oper. Res.*, vol. 135, p. 105456, 2021, doi: 10.1016/j.cor.2021.105456.

[31] H. Sulistiani and A. Tjahyanto, "Comparative Analysis of Feature Selection Method to Predict Customer Loyalty," *IPTEK J. Eng.*, vol. 3, no. 1, p. 1, 2017, doi: 10.12962/joe.v3i1.2257.

[32] N. A. Shaltout, M. El-Hefnawi, A. Rafea, and A. Moustafa, "Information gain as a feature selection method for the efficient classification of influenza based on viral hosts," *Lect. Notes Eng. Comput. Sci.*, vol. 1, no. July, pp. 625–631, 2014.

[33] A. A. Prasetyo and B. Kristianto, "Integration of Iterative Dichotomizer 3 and Boosted Decision Tree to Form Credit Scoring Profile," *Sisforma*, vol. 7, no. 2, p. 58, 2020, doi: 10.24167/sisforma.v7i2.2659.

[34] Dr.J.Arunadevi, S.Ramya, and M. R. Raja, "A study of classification algorithms using Rapidminer," *Int. J. Pure Appl. Math.*, vol. Volume 119, no. 12, pp. 15977–15988, 2018.

e-ISSN: 2614-8404
p-ISSN:2776-3234

The Design of a Monitoring Application System for The Production of Foam Products Using the UML And Waterfall Methods

Henny Yulianti[1], Gatot Tri Pranoto[2] ,
[1]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
[2]IProgram Studi Teknik Informatika, Fakultas Industri Kreatif dan Telematika, Universitas Trilogy
E-mail:hyulia.999@mail.com, gatot.pranoto@trilogi.ac.id

Abstract *The development of information technology, which is followed by a higher level of competition in the foam product industry, encouraging companies to manage their company's resources properly and to plan effective, systematic and mature activities within the company. As a company with a variety of products, the most dominant problem is in the productivity process. Production is the most important part of a manufacturing company, where in carrying out its production activities this company produces based on orders from customers (Job Orders). And the problems that often occur are planning revisions in the midst of production and changing production schedules between groups (lines), delays in production planning in terms of prioritizing planning, and still being done manually in making daily reports. By implementing monitoring, which is the supervision and control of an activity where measurements and evaluations are completed repeatedly from time to time, monitoring is carried out for the purposes of the company and to maintain ongoing management. Monitoring will provide information about the status and trend of production activities towards the company's goals. The solution to this production problem is to build a web-based foam product production monitoring system application using the Waterfall method which is integrated with UML the method used is use case diagrams, activity diagrams, sequence diagrams, class diagrams and component diagrams and software development with PHP and MySQL technology. With Black box testing, it is proven that the design of this foam production monitoring system application can assist the company's foam product production activities in fulfilling customer orders and accurate reports so that it becomes effective and efficient. in improving the productivity and performance of the company.*

Keywords – Production, Monitoring, Foam, UML, Waterfall

I. INTRODUCTION

Good planning and systematic and mature of an activity in the company is a basic characteristic of modern industry. Because basically effective planning of materials, machines, and money will lead to margin gains where this is important in a company.

The problem that is often faced by a company is poor control or supervision in producing goods or services from the company. This causes the production of goods that are not in accordance with what was previously planned. So, the production event which the quantity or quality of production is not in accordance with the planned and applicable standards is very detrimental to the company.

PT Serim Indonesia is a company that makes flexible polyurethane foam materials. Suitable multi-functional foam for any variety of applications which is a fitting material for automobiles, electronics, household goods, shoes, furniture and construction fields. There is sound-absorbing foam, anti-bacterial, reticulated foam, sealable foam, nonyellowing foam and so on.

The problem faced by PT Serim Indonesia is that in its production department, planning revisions often occur in the midst of production and exchange of production schedules between groups (lines), delays in planning required by production in terms of prioritizing planning, and still being done manually in making daily reports. As a company with a variety of products, of course, many problems occur, one of which is in the production section. Production is the most important part of a manufacturing company, where in carrying out its production activities this company produces based on orders from customers (Job Orders).

Based on the problems discussed above, the solution is to build a web-based monitoring information system application that can be accessed online.[1]

Several studies on monitoring applications that have been carried out previously, namely, research on making aircraft health monitoring applications, architectural frameworks using the ISO-13374 standard, Condition Based Maintenance (CBM) using UML diagrams and ARINC 664 standards for transmission of health monitoring data [2]. And research on the use of the Waterfall method to develop a redesigned value-added internal monitoring framework in IT-supported organizational business processes [3]. This research uses SMILE (Self-Monitoring Interactive Learning Evaluation) application design and uses psychological health measurements based on DASS-21, the Waterfall method chosen to develop this application software [4]. Research on the role of blood hypertension monitoring systems,

From the explanations of the four previous studies, in the following the author will explain what distinguishes it from the application of previous research. The application system that will be built is a production monitoring application system using the Waterfall method which is integrated with software development implementing the UML method with a web-based system with PHP and MySQL technology to control foam products at PT. Serim Indonesian. The diagrams that will be used in the UML method are use case diagrams, activity diagrams, sequence diagrams, class diagrams and component diagrams where the UML diagram used is the de facto standard of modeling and analysis language in the software industry [6] and UML is an architectural description language that most widely used and ISO standard [7]. Waterfall was chosen to be used in software development because it is in accordance with the characteristics of software with high quality [8]. This application system will be able to help monitor foam production activities in the company in fulfilling orders from customers quickly and can report them accurately so that it becomes more effective and efficient in producing foam at PT. Serim Indonesian. And this can improve the company's performance and performance.

Based on this, it can be concluded that the problems and objectives and solutions to overcome them, the authors will build a web-based application system for Php and MySQL technology using the Waterfall method which is integrated with UML. This research is entitled "Designing a Monitoring Application System for Foam Product Production Using UML and Waterfall Methods"

II. RESEARCH METHODOLOGI

2.1. UML Software Development Method

In designing the software, the author uses UML (Unified Modeling Language) modeling. Designing a software system, managing complexity is one of the main reasons why you have to make a model [9], modeling helps developers to be able to focus, be able to document, capture the whole system and communicate important aspects of the system being designed [10]

To design the software, the author uses 5 diagrams in UML, namely Use case diagrams, Activity diagrams, Sequence diagrams, Class diagrams and Component Diagrams. By modeling an object-oriented application as follows:[11]

1. Use Case Diagrams
   Describe the interaction between internal systems, external systems, and users. In other words, graphically describes the users who use the system, and with what technique the user relates to the system.
2. Activity Diagrams
   Describe the sequential flow of the activities of a business process or Use Case. Can also be used to model the logic used by the system.
3. Sequence Diagrams
   Describes how objects interact through sending messages (messages) in the execution of a use case or certain operation
4. Class Diagram
   Describe the structure of system objects. Shows the classes that are components of the system, as well as the relationships between classes.
5. Component Diagram
   static. This component diagram shows the organization and dependence of the system/software on pre-existing components

2.2. PHP (Hypertext Preprocessor)

PHP is a scripting language like HTML. In web development, HTML allows dynamic applications to be made that allow data processing and data processing. All the given syntax will be fully executed on the server while only the results are sent to the browser. Then it is a scripted language that is placed on the server and processed on the server [12].

PHP is known as a scripting language, which integrates with HTML tags, is executed on the server, and is used to create dynamic web pages such as Active Server Pages (ASP) or Java Server Pages (JSP). PHP is an open source software.

PHP programs can be activated using an Open Source-based PHP package, namely XAMPP. XAMPP is a PHP package developed by the Open Source community. Xammp provides Apache, MySQL, PHP and phpMyAdmin programs. The advantages of the PHP programming language from other programming languages are as follows:

1. Language PHP programming is a scripting language that does not compile in its use.
2. Web Servers PHP support can be found everywhere from Apache, IIS, Lighttpd, to Xitami with relatively easy configuration.
3. In terms of development, it is easier, because there are many developers who are ready to help in development.
4. In terms of understanding, PHP is the easiest scripting language because it has a lot of references.
5. PHP is an open-source language that can be used on various machines (Linux, Unix, Macintosh, Windows) and can be run at runtime through the console and can also run system commands.

2.3. MySQL

MySQL is a very popular type of database server. MySQL is a type of RDBMS (Relational Database Management System). MySQL supports the PH programming language, a structured query language, because SQL has several rules that have been standardized by an association called ANSI. MySQL is an RDBMS (Relational Database Management System) server. RDBMS is a program that allows database users to create, manage and use data in a relational model. Thus, the tables in the database have a relationship between one table and another. Some of the advantages of MySQL are:[13]

1. Fast, reliable and easy to use. MySQL is three to four times faster than commercial database servers currently available, easy to set up and does not require an expert to administer the MySQL installation.
2. Supported by multiple languages MySQL database server can give error messages in various languages such as Dutch, Portuguese, Spanish, English, French, German, and Italian.
3. Capable of creating very large tables. The maximum size of each table that can be created with MySQL is 4 GB up to a file size that can be handled by the operating system used.
4. Cheaper MySQL is open source and distributed free of charge at no cost to UNIX platforms, OS/2 and Windows Platforms

III. RESULT AND DISCUSSION

3.1. Research Process

In this study the authors used descriptive and action research methods (action research). Descriptive research is research that is intended to collect information about the status of an existing symptom, namely the state of symptoms according to what they were at the time the research was conducted. While the method of action (action research) is research that is used to develop new skills, new approaches, or new knowledge products and to solve problems with direct application in the actual world / field [14].

The research method used by the author is the Waterfall model method because this method takes a systematic and sequential approach. It is called a waterfall because the stages that are passed must wait for the completion of the previous stage and run sequentially. Pictures of the Waterfall model research method can be seen in Figure 1 below:[14]

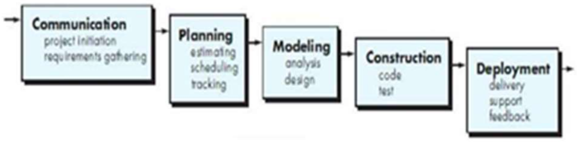


Figure 1. Waterfall Pressman (Pressman, 2015:42) [14]

1.Communication.

The first stage is that the author looks for software needs to be made, by collecting data and information at the research site, namely at PT. Indonesian Series. This process is carried out in several stages as follows:

a. Observations were made at the research site, by observing and seeing how the production process of foam products at the company was.
b. Interviews were conducted with several employees and division heads at the factory where production activities are carried out. This is done to find out, understand the needs and production processes desired by the production department at the company.
c. The author looks for various theories from various sources to support the needs of the software to be designed. And also collect the necessary data from journals, articles and the internet.
d. And analyze the problems faced with the data collected and help define the features and functions of the software according to the needs in monitoring and controlling the production of foam products and according to the conditions and wishes of the company's leadership.

2.Planning (Estimating, Scheduling, Tracking)

The next stage is the planning stage which explains the estimation of technical matters that will be made in the application of monitoring the production of foam products and their uses, conveniences and benefits that can be received by the company and the production department, especially when the monitoring application for the production of foam products is implemented. Then the resources needed to create and design the application system for monitoring the production of this foam product, the final results of the application to be produced, scheduling in making applications to be implemented, and tracking the process of working on the application system.

3. Modeling (Analysis & Design).

This stage is the stage of designing and modeling system architecture that focuses on designing data structures, software architectures, interface displays, and program algorithms. The goal is to better understand the big picture of what will be done in designing the application for this foam product production monitoring system.

4.Construction

This construction stage is the process of translating the design form into a machine-readable code or form/language. After the coding is complete, testing is carried out on the system and also the code that has been created. The goal is to find errors that may occur to be corrected later.

5.Deployment.

The deployment stage is the stage of software implementation to the customer, periodic software maintenance, software repair, software evaluation, and software development based on the feedback provided so that the system can continue to run and develop according to its function.

3.2. Process Design (Design)

In making or designing application monitoring software for the production of foam products, the author uses the UML model method [9][10] which consists of 5 diagrams in UML, namely Use case diagrams, Activity diagrams, Sequence diagrams, Class diagrams and Component Diagrams [11].].

1. Use Case Diagrams.

This research is modeling that will be used to describe the functional requirements of the software that is built by using use case diagrams. The use case diagram consists of 4 (four) actors, namely general admin, warehouse, head of production and marketing. It can be seen in Figure 2 as follows:

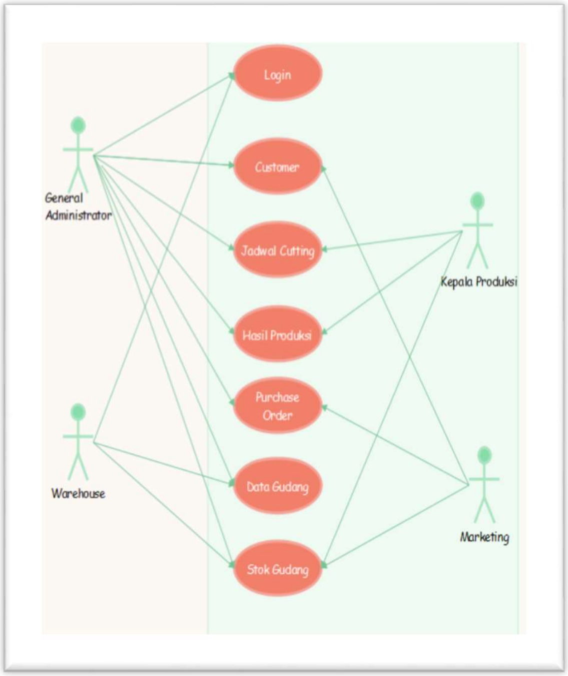


Figure 2. Use Case Diagram

2. Activity Diagrams

Activity diagrams are used to describe a series of flow of activities, used to describe activities that are formed in one operation so that it can also be used for other activities. Figure 3 below illustrates the Activity diagram process for the admin process to process customer data, such as adding new data, changing, and deleting customer data.

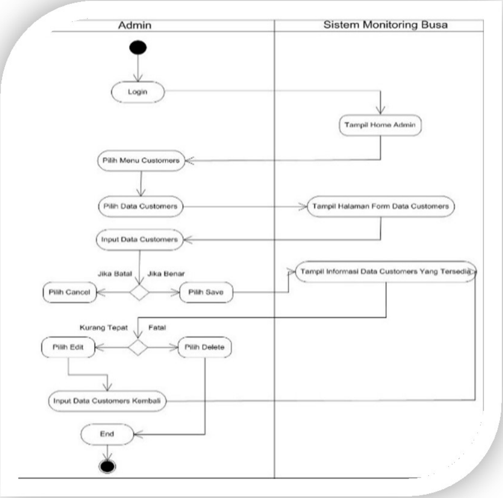


Figure 3. Activity Diagram of the Customer Process

The activity diagram in Figure 4 illustrates the purchase order process, where the admin enters Purchase order data into the system. Starting from the login process by entering the username and password, if incorrect, the user will get an error message from the system, but if it is correct then the user can enter the main page, then the user selects the purchase order data menu and selects the submenu.

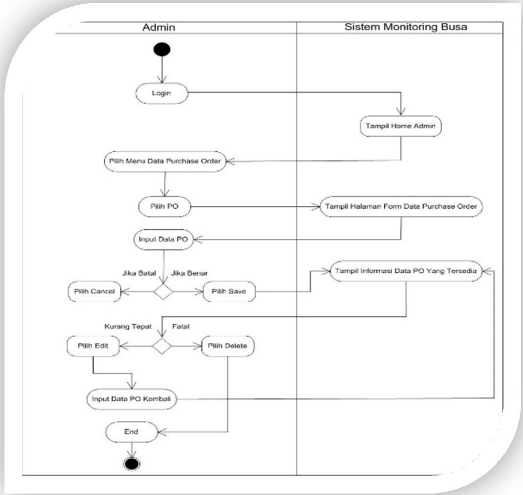


Figure 4. Activity Diagram of the Purchase Order Process

In this Activity diagram, it describes the admin process for processing Production Results data, such as adding new data, changing, and deleting production data. An overview of the Activity diagram of the production data processes as shown in Figure 5 below;

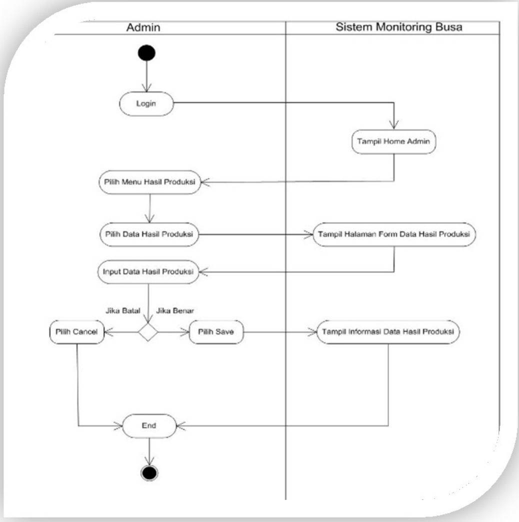

Admin
Sistem Monitoring Busa
Login
Tampil Home Admin
Pilih Menu Hasil Produksi
Pilih Data Hasil Produksi
Tampil Halaman Form Data Hasil Produksi
Input Data Hasil Produksi
Jika Batal
Jika Benar
Pilih Cancel
Pilih Save
Tampil Informasi Data Hasil Produksi
End

Figure 5. Activity Diagram of Production Results

3. Sequence Diagram

Sequence diagrams used to describe a scenario or a series of steps taken in response to an event to produce a certain output, this diagram shows a number of examples of objects and messages placed between objects in a use case diagram.

Sequence Diagram of Goods Data, which describes the flow of goods data processing which is input data starting from logging into the main menu to entering the goods data form to carry out the goods data input process to the process of adding, updating or deleting the goods data. The following is Figure 6, which shows a sequence diagram of the goods data;

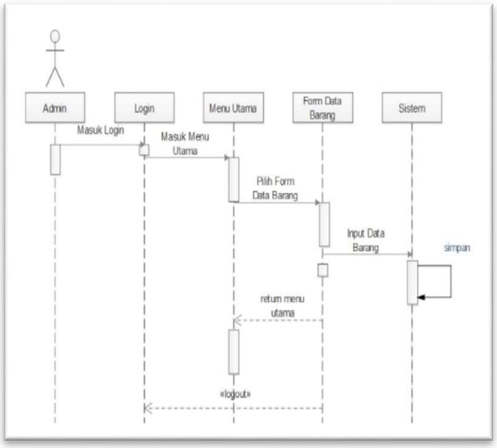


Figure 6. Sequence Diagram of Goods Data

Sequence Diagram of Stock Goods Data, which describes the flow of stock data processing which is input data starting from logging into the main menu to entering the stock item data form to process stock data input to the process for adding, updating or deleting the stock data. The following is Figure 7, which shows a sequence diagram for the stock of goods;

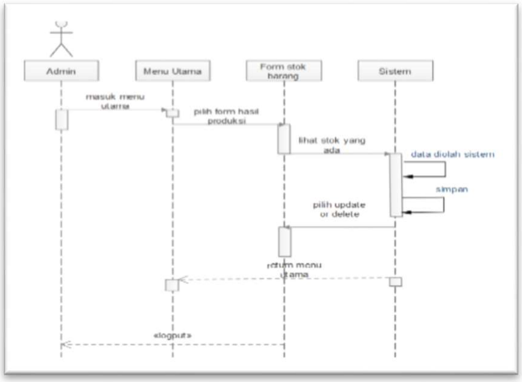


Figure 7. Sequence Diagram of Stock Items

The customer data sequence diagram describes the customer data processing flow which is input data starting from logging into the main menu to entering the customer data form to process customer data input to the process of adding, updating or deleting customer data. The following is Figure 8, which shows the customer sequence diagram;

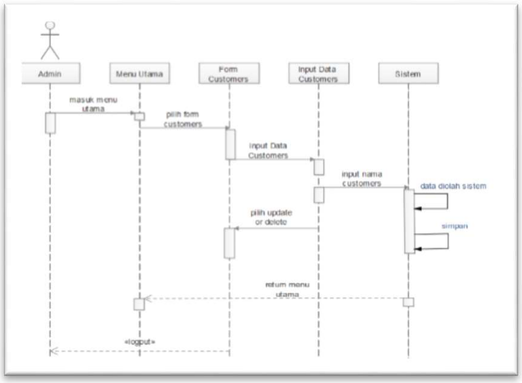


Figure 8. Customer Sequence Diagram

Sequence diagram of the production data describes the process flow of production data processing which is input data starting from logging into the main menu to entering the production data form to carry out the production data input process to the process of adding or deleting the production data. The following is Figure 9, which shows a sequence diagram of the production results;

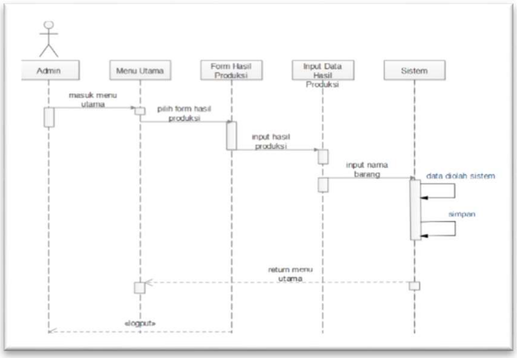


Figure 9. Sequence Diagram of Production Results

4. Class Diagrams

Class Diagram describes the structure of the system in terms of defining the classes that will be created to build a system. Class Diagram is a type of static structure diagram in the Unified Modeling Language (UML) that describes the structure of the system by showing the system classes, their attributes, methods, and the relationships between objects. Figure 10 below, illustrates the static structure of classes in the system foam product production monitoring application

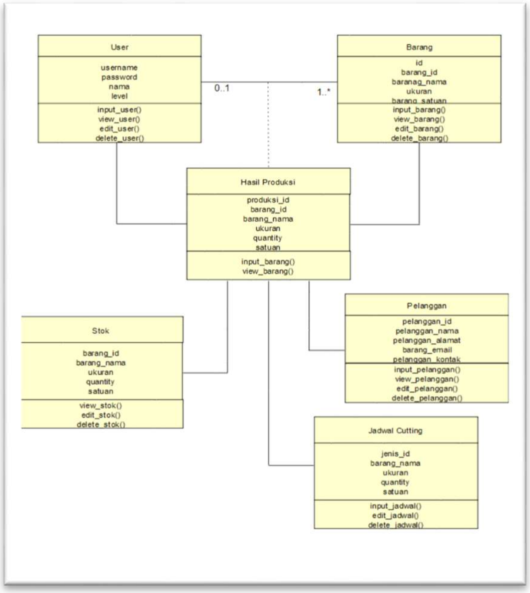


Figure 10. Class Diagram System

5. Component Diagrams

Component diagrams are made to illustrate the structure and dependencies between a collection of components in a system. Figure 11 below is a component diagram of a monitoring application system for foam product production;

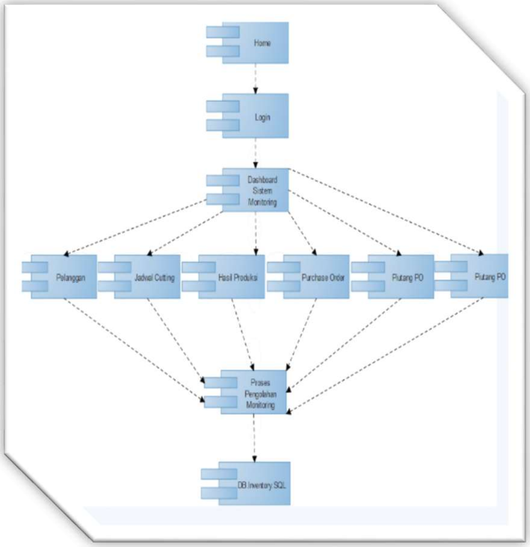


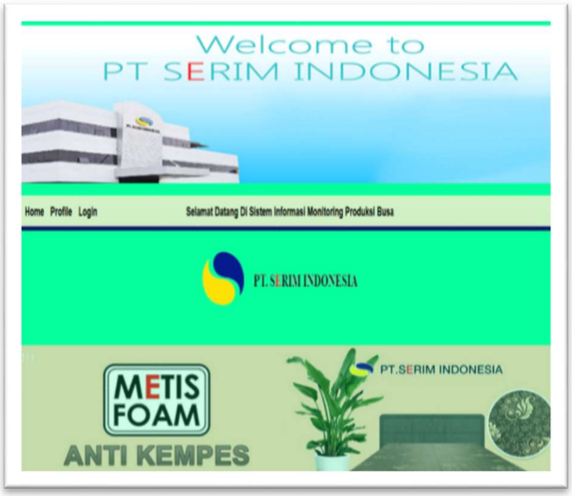
Figure 11. Component Diagram

3.3. Application Design Results

After going through the analysis process and explaining the steps in designing a Web-Based Foam Production Monitoring Information System at PT. Serim Indonesia, the following display of the information system that has been designed:

a. Login Page

The login page is the page used by all system users to enter the main page. In this initial step, the application displays the main page form where the main page form contains information about the initial display / front page



Selamat Datang Di Sistem Informasi Monitoring Produksi Busa

Figure 12. Login Page Display

b. Product Page

Product page is the main page after the profile page form, there is a product menu where the menu displays all the contents of the products in PT. Indonesian Series.

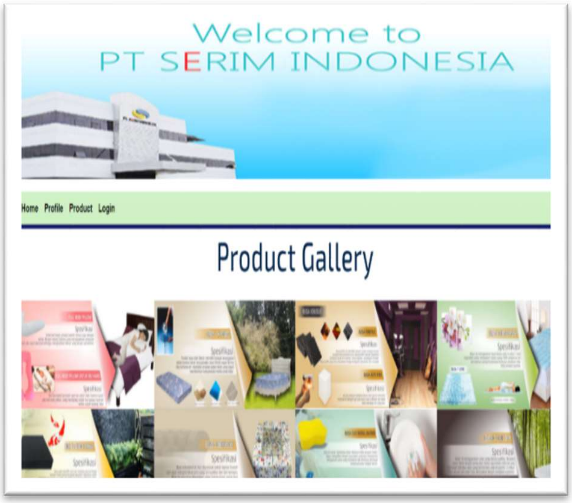


Figure 13. Product Page Display

c. Main Menu Page

The Main Menu will appear after logging in as a user. The display of this form shows what forms can be accessed by the user.

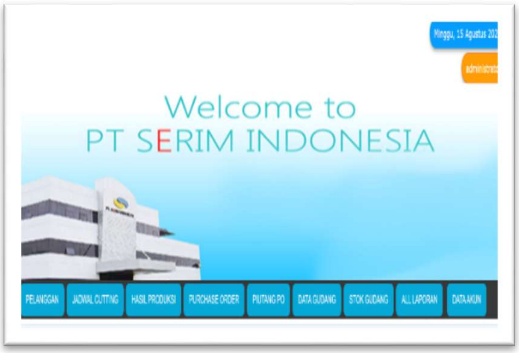
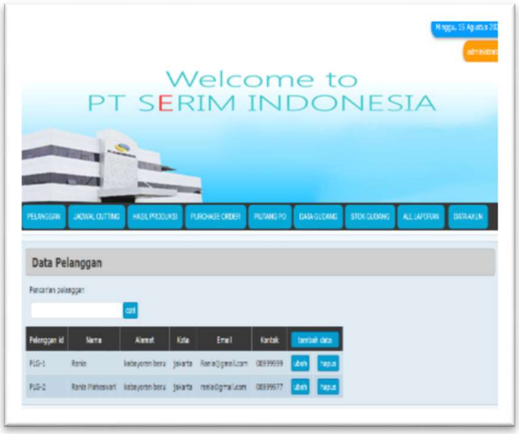


Figure 14. Main Menu Page Display

d. Customer Data Menu Display

This Customer Menu contains all transaction activities consisting of several parts such as: No, Customer Name, Address, City, Email, and Contact



Data Pelanggan

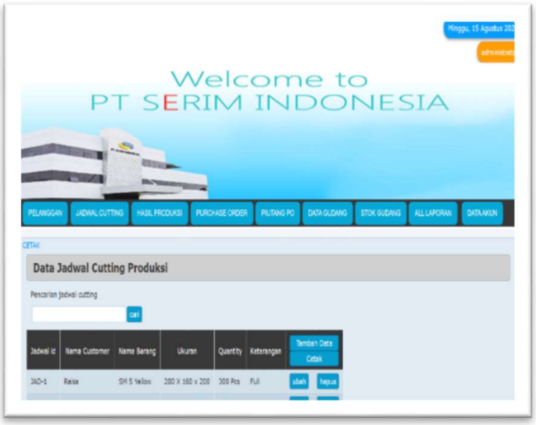
Pencarian pelanggan

Pelanggan id	Nama	Alamat	Kota	Email	Kontak	tambah data
PLG-1	Ranie	kebayoran baru	jakarta	Ranie@gmail.com	08999599	ubah hapus
PLG-2	Ranie Maheswari	kebayoran baru	jakarta	ranie@gmail.com	08999577	ubah hapus

Figure 15. Display of Customer Data Menu

e. Production Cutting Schedule Menu Display

The Production Cutting Schedule Menu contains all transaction activities consisting of several parts such as: No, Customer Name, Item Name, Item Size, Quantity, Part, and Description



Data Jadwal Cutting Produksi

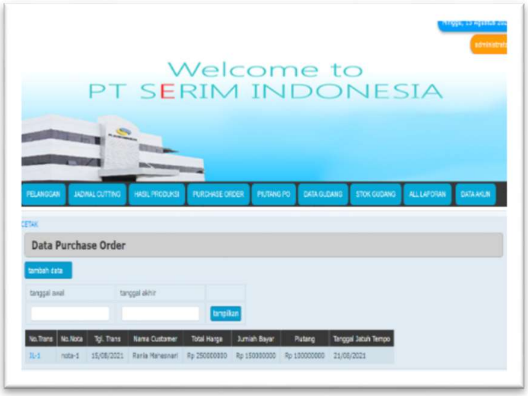
Pencarian jadwal cutting

Jadwal id	Nama Customer	Nama Barang	Ukuran	Quantity	Keterangan	Tamban Data Cetak
JAD-1	Raisa	SM 5 Yellow	200 X 160 x 200	300 Pcs	Full	ubah hapus

Figure 16. Production Cutting Schedule Menu Display

f. Purchase Order Menu Display

This Purchase Order Menu contains all transaction activities consisting of several parts such as: No, No Note, Transaction date, customer name, Amount paid, receivable and due date



Data Purchase Order

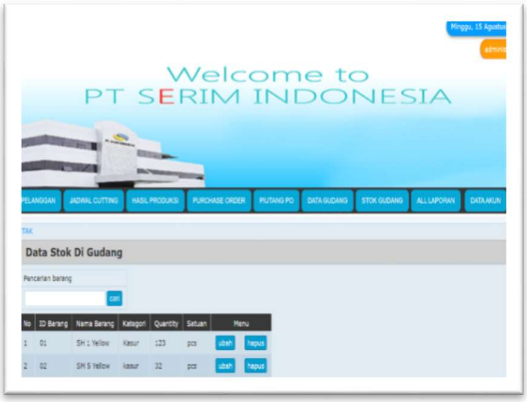
tanggal awal
tanggal akhir

No.Trans	No.Nota	Tgl. Trans	Nama Customer	Total Harga	Jumlah Bayar	Piutang	Tanggal Jatuh Tempo
JL-1	nota-1	15/08/2021	Rania Maheswari	Rp 250000000	Rp 150000000	Rp 100000000	21/08/2021

Figure 17. Display of Purchase Order Menu

g. Warehouse Stock Menu Display

The Warehouse Stock Menu contains all transaction activities consisting of several parts such as: No, Item Name, category, Quantity and unit.



Data Stok Di Gudang

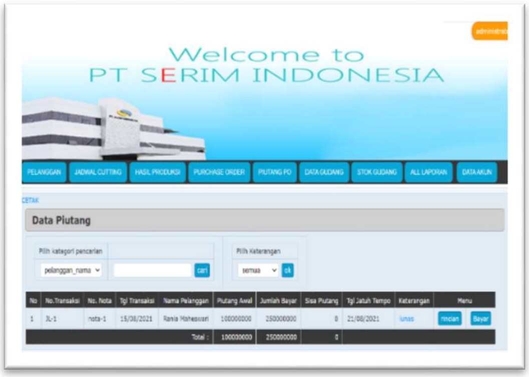
Pencarian barang

No	ID Barang	Nama Barang	Kategori	Quantity	Satuan	Menu
1	01	SH 1 Yellow	Kasur	123	pcs	ubah hapus
2	02	SM 5 Yellow	kasur	32	pcs	ubah hapus

Figure 13. Warehouse Stock Menu Display

h. PO Receivable Menu Display

In the Menu Display this Po Receivable Data contains transaction activities to make payments for the remaining POs that have been previously paid which consist of several parts such as: No, Transaction No, Date, Customer Name, Initial Receivable, Remaining Receivables and Information



Data Piutang

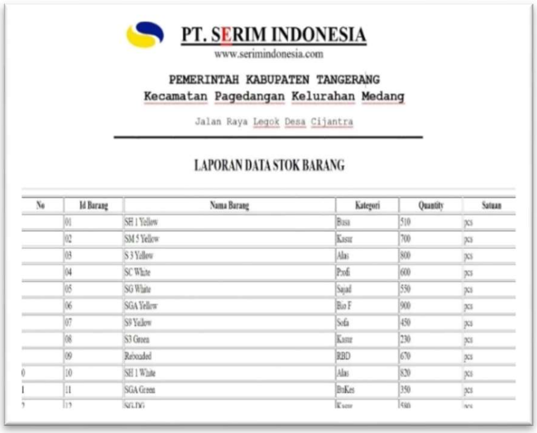
Pilih kategori pencarian: pelanggan_nama
Pilih Keterangan: semua

No	No.Transaksi	No. Nota	Tgl Transaksi	Nama Pelanggan	Piutang Awal	Jumlah Bayar	Sisa Piutang	Tgl Jatuh Tempo	Keterangan	Menu
1	JL-1	nota-1	15/08/2021	Rania Maheswari	100000000	250000000	0	21/08/2021	lunas	rincian Bayar
				Total :	100000000	250000000	0			

Figure 14. Report Menu Page Display

I. Report Print View

In the print view, this report contains the process of report data to be printed.



PT. SERIM INDONESIA
www.serimindonesia.com

PEMERINTAH KABUPATEN TANGERANG
Kecamatan Pagedangan Kelurahan Medang
Jalan Raya Legok Desa Cijantra

LAPORAN DATA STOK BARANG

No	Id Barang	Nama Barang	Kategori	Quantity	Satuan
	01	SH 1 Yellow	Busa	510	pcs
	02	SM 5 Yellow	Kasur	700	pcs
	03	S 3 Yellow	Alas	800	pcs
	04	SC White	Profi	600	pcs
	05	SG White	Sajad	550	pcs
	06	SGA Yellow	Bio F	900	pcs
	07	S9 Yellow	Sofa	450	pcs
	08	S3 Green	Kasur	230	pcs
	09	Rebounded	RBD	670	pcs
0	10	SH 1 White	Alas	820	pcs
1	11	SGA Green	BnKes	350	pcs
2	12	SGDG	Kasur	580	pcs

Figure 15. Print Report Page Display

3.4. Black Box Test

At the Black box testing stage of the online Foam Product Production Monitoring System application, testing is carried out by running all the functions and features available from this application and then seeing whether the results of these functions are as expected, [15]

Testing is carried out, using the assumption of not knowing the internal structure of the program (black box). Concentrate on finding conditions where the program does not run according to specifications (functional) using specifications for test data [16].

This test is carried out after the system is made by testing all the existing buttons. This test ensures whether the process carried out produces output that is appropriate or not in accordance with the design.[17]

Table 1. BLACK BOX TEST

No	Things to Test	Results obtained	Results
1	Page Main	**a.** Go back to the main page. **b.** Go to the profile page. **c.** Go to the product page. **d.** Go to the login page.	In accordance
2	Menu Main	a. Go to the main menu page. b. Go to customer data page c. Go to the production cutting schedule page. d. Go to the production page. e. Go to the purchase order page. f. Go to the item data page. g. Go to stock data page h. Go to accounts receivable page. i. Go to Report Page. j. Go to the user data page. k. Log out of the system.	In accordance
3	Production Cutting Schedule Form	a. Displays additional production cutting schedule data. b. Displays in detail the available cutting schedule. c. Displays changes to the production cutting schedule data process. d. Displaying deletions in production cutting schedule data	In accordance

4	Production Result Form	a. Displays the addition of production data. b. Displays print for production results. c. Displays changes in the production data process. d. Displays deletions on production data.	In accordance
5	Purchase Order Form	a. Displays additions to the purchase order data. b. Display print on purchase order data c. Displays changes to the purchase order data process. d. Displaying deletions on purchase order data	In accordance
6	Customer Data Form	a. Displays additions to customer data. b. Displays changes to the customer data process. c. Displays deletion on customer data.	In accordance
7	Item Data Form	a. Displays additions to item data. b. Displays changes to the process of goods data. c. Displays deletion on item data.	In accordance
8	Stock Data Form	a. Displays additions to stock data. b. Displays changes to the stock data process. c. Displays deletions on stock data.	In accordance
9	From All Report	a. Displays the Cutting Schedule Report. b. Displaying Production Results Report. c. Displays Purchase Order Report. d. Displaying PO Accounts Receivable Report. e. Displays the Cutting Schedule Report.	In accordance
10	Account Data Form	a. Displays admin data storage When making changes to user data. b. Displays reset on user data When making changes to user data. c. Displays cancel on user data.	In accordance

♦ Based on the Black Box testing, the login and menus carried out in this application are to check whether the functionality of the menu and login form has been running well and testing for user-accessible menus in the foam production monitoring application is functioning properly.

IV. CONCLUSION

Based on the results of research, design and testing that has been carried out on the application of monitoring the production of foam products PT. Serim Indonesia with the stages of the process that has been carried out starting from planning, analysis, design, coding, implementation and testing. Black box, the authors draw the following conclusions:

1. With advances in information technology today, the development of a Web-based Foam Product Production Monitoring System using the Waterfall method which is integrated with the UML software development method is going well. And it is proven that applications designed with PHP and MySQL technology can be compatible with both online.
2. The results of the Blackbox test, indicate that all functionality in the application has been running well in accordance with online planning and design. Where it is proven that the Foam Product Production Monitoring Application that has been designed can control foam product production activities in fulfilling customer orders quickly and with accurate reports. So that it can increase the production activities of foam products and the company's performance.

V. REFERENCES

[1] X. Zhang, B. Zheng and L. Pan, "*Using Virtual Reality Technology to Visualize Management of College Assets in the Internet of Things Environment*," in IEEE Access, vol. 8, pp. 157089-157102,2020,doi:10.1109/ACCESS.2020.3019836.

[2] AK Jha, S. Nayak and NK Veerabhadrappa, "*An Architecture for Performing Real Time Integrated Health Monitoring of Aircraft Systems Using Avionics Big Data*," 2nd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), 2017, pp. 1-5, Doi:10.1109/CSITSS.2017.8447679

[3] R. Van Hillo and H. Weigand, "*Continuous Auditing & Continuous Monitoring: Continuous Value*?" IEEE Tenth International Conference on Research Challenges in Information Science (RCIS), 2016, pp. 1-11, Doi:10.109/RCIS.2016.7549279.

[4] WA Syafei et al., "*SMILE (Self-Monitoring Interactive Learning Evaluation) for Indonesian University Students*," International Biomedical Instrumentation and Technology Conference (IBITeC), 2019, pp.12-16, Doi: 10.1109/ IBITeC465 97.2019.9091726

[5] AV Demidov, DV Papshev and LY Krivonogov, "*Principles of Construction, Structure and Features of the ECG and Blood Pressure Monitoring System,*" 2020 Moscow Workshop on Electronic and Networking Technologies (MWENT), 2020, pp. 1-5, Doi:10.109/MWENT47943.2020.9067390.

[6] Y. Yang, W. Ke, J. Yang and X. Li, "*Integrating UML With Service Refinement for Requirements Modeling and Analysis*," in IEEE Access, vol. 7, pp.

11599-11612, 2019, Doi: 10.1109/ACCESS.2019. 2892082.

[7] J. MaIm, F. Ciccozzi, J. Gustafsson, B. Lisper and J. Skoog, "*Static Flow Analysis of the Action Language for Foundational UML*," IEEE 23rd International Conference on Emerging Technologies & Factory Automation (ETFA), 2018, pp. 161-168, Doi:10.1109/ETFA.2018.8502620.

[8] S. Balaji and MA Obaidy, "*Project characteristics used for methodology selection to develop the software project*," International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016, pp. 3570-3573, Doi: 10.109/ICEEOT.2016.7755370.

[9] A. Hafeez Khan, S. Hyder Abbas Musavi, A. Rehman and A. Shaikh. "*Ontology-Based Finite Satisfiability of UML Class Model*", IEEE Access, vol.6, pp.3040-3050. Doi:10.1109/ACCESS.2017 2786781.

[10] Samuel Szoniecky, "*Graphical Specifications for Modeling Existence*", in Ecosystems Knowledge: Modeling and Analysis Method for Information and Communication, Wiley, pp.89-115.Doi:10. 1002/978 1119388777.ch4.

[11] L. Ordinez, G. Eggly, M. Micheletto and R. Santos, "*Using UML for Learning How to Design and Model Cyber-Physical Systems*". IEEE Revista Iberoamericana de Tecnologias del Aprendizaje (RITA), vol.15, no.1, pp.50-60. Doi:10. 1109/ RITA. 2020.2978416.

[12] SI Adam and S. Andolo, "*A New PHP Web Application Development Framework Based on MVC Architectural Pattern and Ajax Technology*". 1st International Conference on Cybernetics and Intelligent Systems (ICORIS), pp. 45-50, Doi:10.1109/ ICORIS. 2019.8874912.

[13] MM Eyada, W. Saber, MM El Genidy and F. Amer, "*Performance Evaluation of IoT Data Management Using MongoDB Versus MySQL Databases in Different Cloud Environments*". IEEE Access, vol.8, pp.110656-110668.Doi:10.1109/ACCESS.2020. 3002164.

[14] Pressman R, S, "*Software Engineering*", Publisher Andi Yogyakarta, (2015) Page 42.

[15] TH Chang, J. Larson, LT Watson and TCH Lux, "*Managing Computationally Expensive Blackbox Multiobjective Optimization Problems with Libensemble*", Spring Simulation Conference (SpringSim), (2020) pp.1-12. Doi:10.22360/ SpringSim. 2020.HPC.001.

[16] J.Pan, "*Blackbox Trojanising of Deep Learning Models: Using Non-Intrusive Network Structure and Binary Alterations*", IEEE Region 10 Conference (TENCON), (2020), pp.891-896. Doi:10.1109 /TENCON50793.2020.9293933.

[17] R Guidotti, A Monneale, F Giannotti, D Pedreschi, S Ruggieri F Turini, (2019)," *Factual and Counterfactual Explanations for Black Box Decision Making",* IEEE Intelligent Systems, Vol 34, No6, pp,14-23. Doi:10.1109/MIS.2019.2957223

Application of Feature Selection for Identification of Cucumber Leaf Diseases (*Cucumis sativa L.*)

Lalitya Nindita Sahenda[1*)], Ahmad Aris Ubaidillah[2], Zilvanhisna Emka Fitri[3], Abdul Madjid[4], Arizal Mujibtamala Nanda Imron[5]

[1]Program Studi Teknik Komputer, Jurusan Teknologi Informasi, Politeknik Negeri Jember
[2,3]Program Studi Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Jember
[4]Program Studi Budidaya Tanaman Perkebunan, Jurusan Produksi Pertanian, Politeknik Negeri Jember
[4]Program Studi Teknik Elektro, Fakultas Teknik, Universitas Jember
Email: [1]lalitya.ns@polije.ac.id, [2]arisubay06@gmail.com, [3]zilvanhisnaef@polije.ac.id, [4]abdul_madjid@polije.ac.id, [5]arizal.tamala@unej.ac.id

Abstract − According to data from BPS Kabupaten Jember, the amount of cucumber production fluctuated from 2013 to 2017. Some literature also mentions that one of the causes of the amount of cucumber production is disease attacks on these plants. Most of the cucumber plant diseases found in the leaf area such as downy mildew and powdery mildew which are both caused by fungi (fungal diseases). So far, farmers check cucumber plant diseases manually, so there is a lack of accuracy in determining cucumber plant diseases. To help farmers, a computer vision system that is able to identify cucumber diseases automatically will have an impact on the speed and accuracy of handling cucumber plant diseases. This research used 90 training data consisting of 30 healthy leaf data, 30 powdery mildew leaf data and 30 downy mildew leaf data. while for the test data as many as 30 data consisting of 10 data in each class. To get suitable parameters, a feature selection process is carried out on color features and texture features so that suitable parameters are obtained, namely: red color features, texture features consisting of contrast, Inverse Different Moment (IDM) and correlation. The K-Nearest Neighbor classification method is able to classify diseases on cucumber leaves (Cucumis sativa L.) with a training accuracy of 90% and a test accuracy of 76.67% using a variation of the value of K = 7.

Keywords – Identification, Cucumber Leaf, Disease, Color Feature, GLCM, KNN

I. INTRODUCTION

Cucumber (Cucumis sativa L.) is one of the most widely consumed fruit vegetables and the growing conditions are very flexible, because it can grow in both highlands and lowlands [1]. In Indonesia, the amount of cucumber production fluctuated (instability) from 2013 to 2017. Total cucumber production was 9.97 tons/ha in 2013, then decreased to 9.84 tons/ha in 2014 and increased to 10.27 tons/ha in 2015. In 2016, production decreased again to 10.19 tons/ha and increased again to 10.67 tons/ha in 2017 [2]. One of the causes of the fluctuating amount of cucumber production is the attack of pests and diseases on the plant. Symptoms of disease in plants can be seen from the plant body parts, such as leaves, fruit, stems and roots. Most of the cucumber plant diseases that are found in the leaf area such as antracnose lesion [3], downy mildew and powdery mildew are usually caused by fungi (fungal disease) [4].

So far, farmers have checked for cucumber plant diseases, which are still done manually based on the experience of farmers. This determination certainly has weaknesses, one of which is the lack of accuracy in determining cucumber plant diseases. To help farmers, a cucumber leaf disease identification system was created using computer vision. Computer vision is a branch of science where a system utilizes digital image processing techniques which are then analyzed using artificial intelligence.

Some studies that become the reference of this research are cucumber disease detection using diagnosis of diseases of cucumber through extracting three characteristic values of shape, texture and color [5][3], Then the research was developed by adding image smoothing and edge detection so that the segmentation results were better[6]. The feature extraction technique used is first-order statistical features and second-order statistical features such as Gray Level Co-Occurrence Matrix (GLCM) to get an accuracy of 80.45% [7], then developed the introduction of cucumber disease using sparse representation classification with an accuracy rate of 85.7% using the classification method used in this study is K-Nearest Neighbor [8]. The KNN method is also able to classify other research objects such as the classification of platelets on peripheral blood smears with an accuracy of 83.67% [9], on the classification of tuberculosis bacteria with an accuracy of 94.92% [10], Classification of white blood cell abnormalities based on shape features (area, perimeter, metric and compactness) with an accuracy rate of 94.3% [11] and classification of bacteria that cause ARI with an accuracy of 91.67% using a variation of the value of K = 3.5 and 7 [12].

To classify data, the KNN method uses the closest distance to the object so that the method is often known as lazy learning. The basic principle of KNN is to find the value of K where the value of K is the closest amount of data that will determine the classification results and to calculate the closest distance using Euclidean distance calculations. Based on this description, the K-Nearest Neighbor (KNN) method is able to classify cucumber leaf diseases with a good level of accuracy.

II. RESEARCH METHODOLOGY

This research was conducted in a cucumber field in Lumajang Regency, East Java. Data collection is done with the help of direct sunlight, so the resulting image is very bright and does not cause shadows (noise). This will affect the results in the image processing process. The system flow consists of starting from the image of cucumber leaves, then the image is processed using digital image processing techniques, the result of feature extraction becomes the input of the KNN classification method which will be classified into 3 classes as shown in Figure 1.

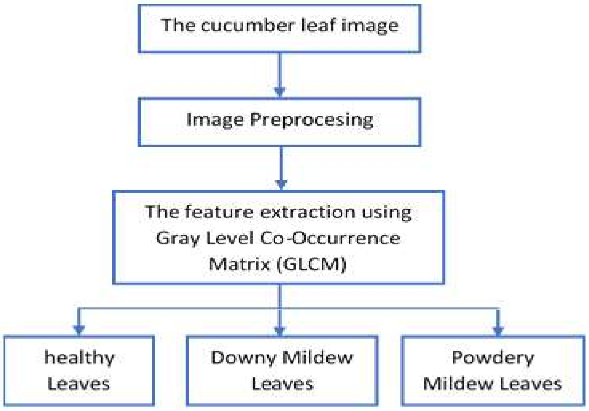
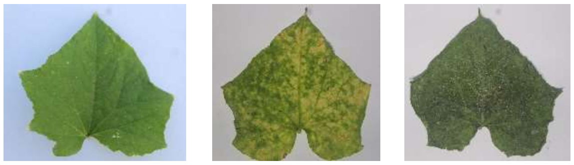


Figure 1. The Cucumber Leaf Identification System Diagram

A. *The Cucumber Leaf Images Data*

Image data was taken using a Canon EOS 1100D camera with a camera resolution of 12 MP, using a tripod and studio light box. The data is divided into 3 classes, namely healthy leaf images, downy mildew leaf images and powdery mildew leaf images as shown in Figure 2. This research used 90 training data consisting of 30 healthy leaf data, 30 powdery mildew leaf data and 30 downy mildew leaf data. while for the test data as many as 30 data consisting of 10 data in each class.

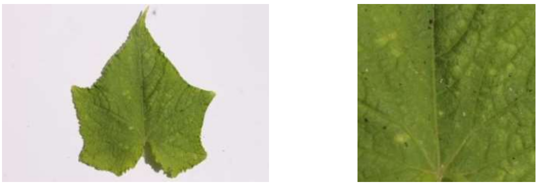


(a) **(b)** **(c)**

Figure 2. Image of cucumber leaves (a) healthy, (b) downy mildew and (c) powdery mildew

B. *Image Processing*

The image processing process is the initial stage of the image processing process which aims to improve image quality and the data normalization process. This process begins with a cropping process whose aim is to obtain a smaller image size to reduce the computational load [13]. The initial image size of 4272 x 2848 pixels is cropped to 1001 x 1001 pixels as shown in Figure 3.



(a) **(b)**

Figure 3. Image (a) original and (b) after cropping process

C. *Feature Extraction*

Feature extraction aims to extract the unique characteristics of the cucumber leaf image. Before performing feature extraction, the original image is an RGB color space image, then the process of splitting the color components into red, green and blue as shown in Figure 4.

In this study, two features were used, namely color and texture. The color features used are red, green and blue color features, while the texture features used are texture features from the gray level co-occurrence matrix (GLCM) method.

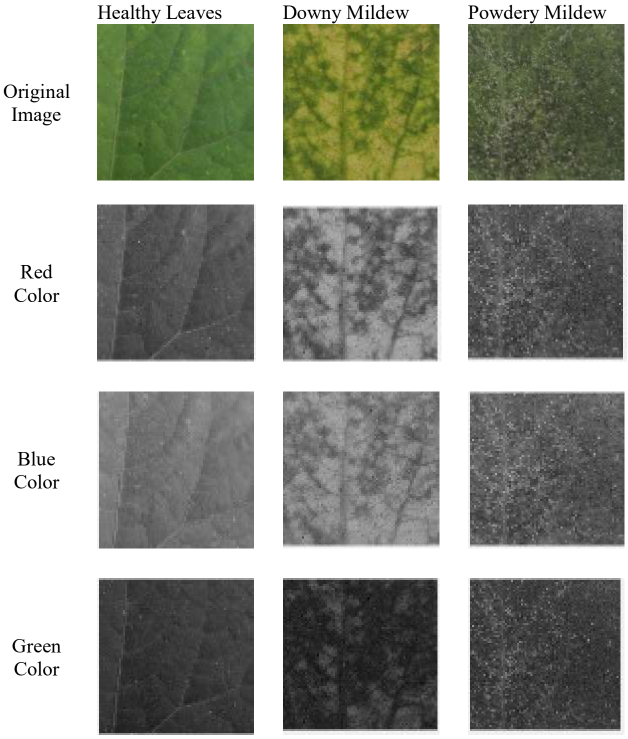


Figure 4. Image of the RGB components based on each classes

GLCM is a gray degree matrix that represents the frequency of occurrence of two pixels with a certain intensity in a distance d and a certain angle direction θ. Therefore, the matrix provides information that differs from the difference in distance between pixels [14]. The angles used are 0°, 30°, 45°, 90° and 135°. The formula equation used is as follows [15]:

$$\mathrm{ASM} = \sum_{i=1}^{L}\sum_{j=1}^{L}(\mathrm{GLCM}(i,j))^2 \qquad (1)$$

$$\mathrm{contrast} = \sum_{i}^{L}\sum_{j}^{L}|i-j|^2\mathrm{GLCM}(i,j) \qquad (2)$$

$$\mathrm{IDM} = \sum_{i=1}^{L}\sum_{j=1}^{L}\frac{(\mathrm{GLCM}(i,j))^2}{1+(i-j)^2} \qquad (3)$$

$$\mathrm{Entropy} = -\sum_{i=1}^{L}\sum_{j=1}^{L}(\mathrm{GLCM}(i,j))\log(\mathrm{GLCM}(i,j)) \qquad (4)$$

$$\mathrm{Correlation} = \sum_{i=1}^{L}\sum_{j=1}^{L}\frac{(i-\mu i')(i-\mu j')(\mathrm{GLCM}(i,j))}{\sigma i\sigma j} \qquad (5)$$

D. *K-Nearest Neighbor Classification*

One of the easy classification methods is the K-Nearest Neighbor classification method. This method has the basic principle of looking for a constant K value where the K value is the number of closest distances that affect the classification results. The calculation of distance using Euclidean distance calculation with the equation [12] :

$$d_{(xi,xj)} = \sqrt{\sum_{r=1}^{n}(Xir - Xij)^2} \qquad (6)$$

III. Results And Discussion

In this research, the results of the cropping process are taken for the color features of each component of the RGB color space as shown in Figure 4. The image shows that in the blue and green images there is no significant difference in the classes: healthy leaves, downy mildew and powdery mildew. On the other hand, the red image shows a significant difference, especially in the powdery mildew leaf class, so that the red image is the best in representing the textures of the three classes. In the image of the red component, the gray level values on the downy mildew and powdery mildew leaves are clearly visible so that the red image becomes the input for texture feature extraction based on the Gray Level Co-Occurrence Matrix (GLCM) value.

A. Color Feature Extraction

In this research, color features were taken so that each class was as shown in Table 1. The table shows that there is a closeness of values for the Blue features of the Healthy leaf class and the downy mildew leaf class. Whereas in the Green feature, there is also a closeness of values in the Healthy leaf class and the powdery mildew leaf class, so that the color feature used is the red image feature.

Table 1. The Average Value of The RGB Color Feature in Each Class

Class	*Red*	*Green*	*Blue*
Healthy	89.29	110.83	39.80
Downy mildew	118.84	125.34	39.64
Powdery mildew	98.27	111.81	57.44

B. Texture Feature Extraction

This research also takes texture features using GLCM texture features. To take the GLCM feature, a red image is used as the input image, as previously explained that the red image best represents the texture in the three classes. The average value of texture features is shown in Table 3, where the features used are Angular Second Moment (ASM), Contrast, Inverse Different Moment (IDM), Entropy and Correlation. Table 3 shows that there is a closeness of values for the ASM features of the three classes, while for the entropy features there is also a closeness of values for the downy mildew and powdery mildew class.

Table 2. The Average Value of GLCM Features in Each Class

Class	ASM	Contrast	IDM	Entropy	Correlation
Healthy	0.0068	4.8496	0.6109	5.6273	0.0076
Downy mildew	0.0026	5.5866	0.5265	6.4159	0.0024
Powdery mildew	0.0030	21.5099	0.3883	6.6609	0.0044

C. Feature Selection

Based on the discussion on color feature extraction and texture feature extraction, it was found that some features have close values and this will affect the classification process. There will be errors in the classification process due to the proximity of these values, so the researcher conducts a feature selection process and the features used are red color features, contrast texture features, IDM, and Correlation. An example of selecting a red feature is shown in Figure 5.

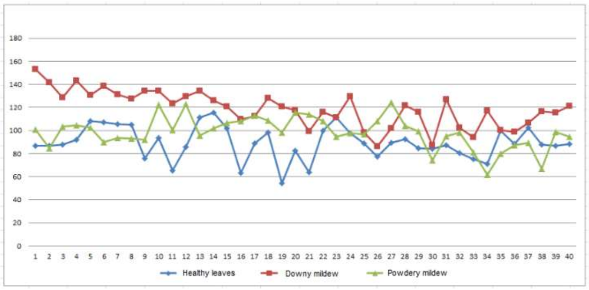
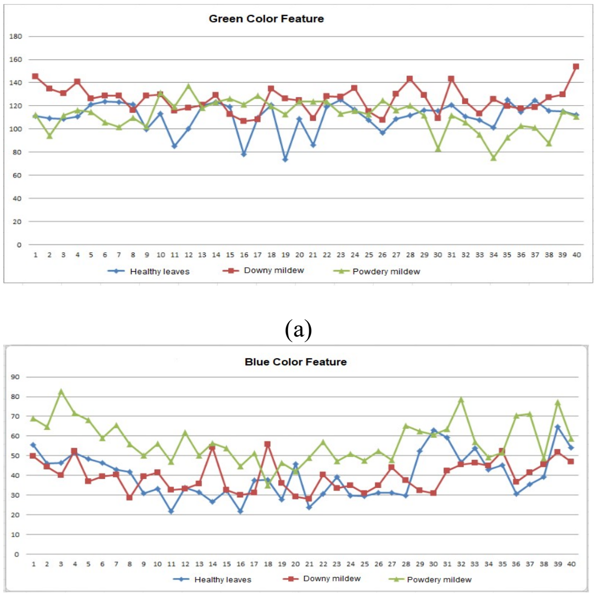


Figure 5. Graph of data distribution in each class on the red feature

Figure 5 shows that the blue color describes the distribution of healthy leaf data, the red color describes the distribution of downy mildew data and the green color describes the distribution of powdery mildew data. The graph will be different when compared to the green and blue color features as shown in Figure 6.



(a)

(b)

Figure 6. Graph of data distribution in each class on (a) green and (b) blue color features

On the graph (Figure 6), there is still a lot of data that occurs in the values between the three classes so that the graphs of the three classes intersect. While the GLCM features a graph depicting the value of each feature in the three classes (Figure 7).

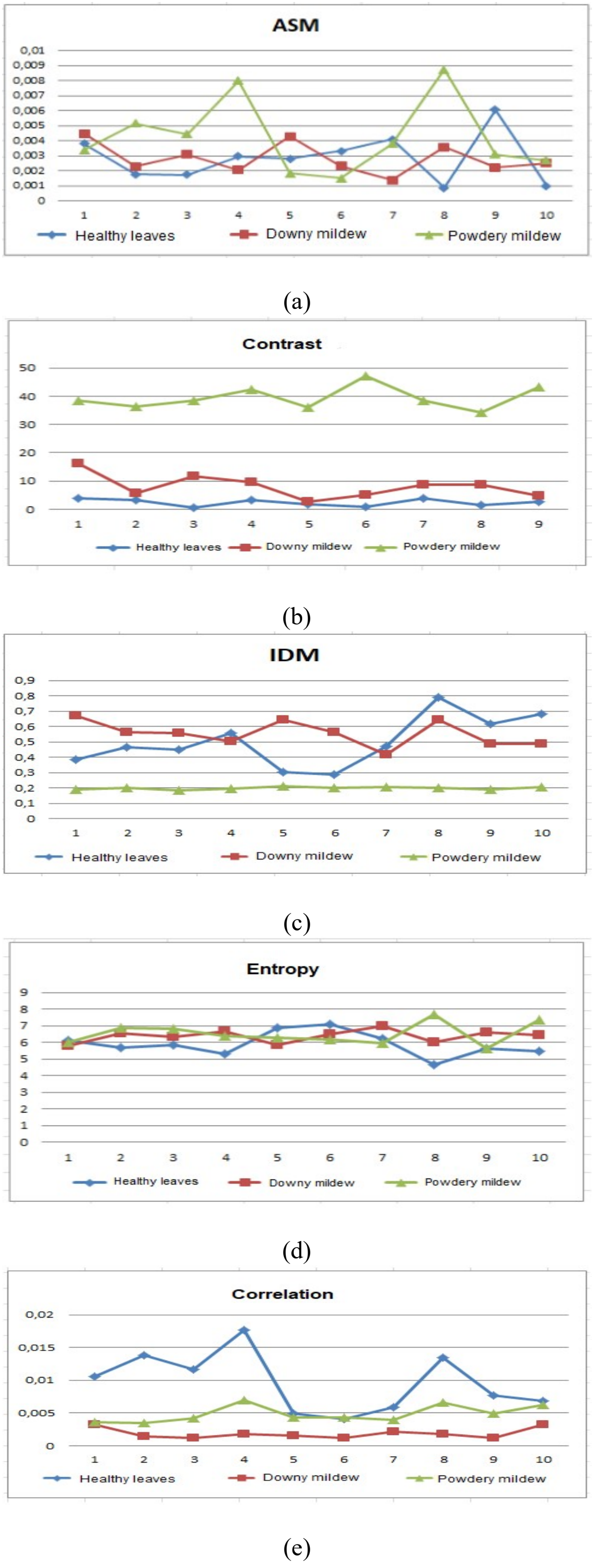


Figure 7. Graph of data distribution in each class on (a) ASM, (b) contrast, (c) IDM, (d) entropy and (e) correlation features

On Figure 7 shows that in features (a) ASM and (d) entropy features, there are many data that have values that are tangent to each other in the three classes. this affects the accuracy of the system in classifying data so that the two features become features that will reduce the system in classifying the three classes. Whereas in features (b) contrast, (c) IDM and (e) correlation, there is a difference in value even though there are features whose values intersect on healthy leaf data with downy mildew data or on healthy leaf data with powdery mildew data. Based on the results of feature selection, this study used 4 features consisting of color features and texture features, namely red color features, texture features (contrast, IDM and correlation) as shown in Table 3.

Table 3. The Examples of Feature Selection Results for Each Class

Feature	**Class**			
	Red	**Contrast**	**IDM**	**Correlation**
Healthy Leaves	86.71	8.7433	0.3836	0.0105
	86.94	5.2779	0.4650	0.0138
	87.96	5.4021	0.4503	0.0116
Downy mildew	153.04	0.6716	0.0030	2.3340
	141.83	0.5647	0.0014	4.5463
	128.70	0.5571	0.0027	4.3620
Powdery Mildew	93.70	0.4850	0.0069	6.9519
	92.82	0.2257	0.0021	7.5547
	92.22	0.6759	0.0021	5.4755

D. K-Nearest Neighbor Classification

The number of training data is 90 cucumber leaf image data consisting of 30 healthy leaf data, 30 powdery mildew leaf data, 30 downy mildew leaf data. While 30 test data consisting of 10 data in each class. The results of the accuracy of the KNN classification method. Based on Table 4, the highest accuracy of system training is 100% at the variation of the value of K = 1. However, the accuracy of the test is 66.77%, so that judging from the results of the highest test accuracy, it is 76.67% with a training accuracy of 90% at the variation of the value of K = 7.

Table 4. The Percentage of System Training and Testing Accuracy

K Value	**Training Accuracy (%)**	**Testing Accuracy (%)**
1	100	66.67
3	96.67	73.33
5	93.33	70
7	90	76.67
9	88.89	73.33
11	90	70
13	90	66.67
15	87.78	66.67

The calculation of system accuracy is obtained based on ROC calculations with a confusion matrix as shown in Table 5 for the training process and Table 6 for the testing process.

Table 5. The Confusion Matrix in The Training Process

Classification Results			**Target**
Healthy	Downy Mildew	Powdery Mildew	
24	0	6	Healthy
0	30	0	Downy Mildew
3	0	27	Powdery Mildew

$$Accuracy\ of\ training = \frac{24+30+27}{24+30+27+3+6} x100\%$$

$$= 90\%$$

Table 6. The Confusion Matrix in The Testing Process

Classification Results			Target
Healthy	Downy Mildew	Powdery Mildew	
8	0	2	Healthy
0	10	0	Downy Mildew
4	1	5	Powdery Mildew

$$Accuracy\ of testing = \frac{8+10+5}{8+10+5+2+4+1} x100\%$$

$$= 76.67\%$$

Based on the results of these calculations, it can be seen that there is a significant difference in accuracy in the training and testing process. The system training accuracy rate is 90% while the system testing accuracy is 76.67%. This can happen due to the lack of data used or there is a significant difference in the value of the training data and testing data.

IV. Conclusion

The K-Nearest Neighbor (KNN) classification method is able to classify diseases on cucumber leaves (Cucumis sativa L.) with a training accuracy of 90% and a test accuracy of 76.67% using a variation of the value of K = 7. The lack of data also affects the classification results because the system has limitations in recognizing patterns from healthy leaf classes, downy mildew and powdery mildew. In addition, this research must also compare other classification methods.

References

[1] 1. KL, A., Napitupulu, M., & Jannah, N. (2015). RESPON TANAMAN MENTIMUN (Cucumis sativus L.) TERHADAP JENIS POC DAN KONSENTRASI YANG BERBEDA. *AGRIFOR*, *XIV*(1), 15–26.

[2] BPS Kabupaten Jember. (2020). *Kecamatan Arjasa Dalam Angka Tahun 2020.*

[3] Pixia, D., & Xiangdong, W. (2013). Recognition of Greenhouse Cucumber Disease Based on Image Processing Technology. *Open Journal of Applied Sciences*, *03*(01), 27–31. https://doi.org/10.4236/ojapps.2013.31b006

[4] Pawar, P., Turkar, V., & Patil, P. (2016). Cucumber disease detection using artificial neural network. *Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016*, *2016*. https://doi.org/10.1109/INVENTIVE.2016.7830151

[5] Wei, Y., Chang, R., Wang, Y., Liu, H., Du, Y., Xu, J., & Yang, L. (2012). A study of image processing on identifying cucumber disease. *IFIP Advances in Information and Communication Technology*, *370 AICT*(PART 3), 201–209. https://doi.org/10.1007/978-3-642-27275-2_22

[6] Jiannan, J., & Haiyan, J. (2013). Recognition for cucumber disease based on leaf spot shape and neural network. *Transactions of the Chinese Society of Agricultural Engineering*, *29*(1).

[7] Kaur, S., Pandey, S., & Goel, S. (2019). Plants Disease Identification and Classification Through Leaf Images: A Survey. *Archives of Computational Methods in Engineering*, *26*(2), 507–530. https://doi.org/10.1007/s11831-018-9255-6

[8] Khan, M. A., Akram, T., Sharif, M., Javed, K., Raza, M., & Saba, T. (2020). An automated system for cucumber leaf diseased spot detection and classification using improved saliency method and deep features selection. *Multimedia Tools and Applications*, *79*(25–26), 18627–18656. https://doi.org/10.1007/s11042-020-08726-8

[9] Fitri, Z. E., Purnama, I. K. E., Pramunanto, E., & Purnomo, M. H. (2017). A comparison of platelets classification from digitalization microscopic peripheral blood smear. *2017 International Seminar on Intelligent Technology and Its Application: Strengthening the Link Between University Research and Industry to Support ASEAN Energy Sector, ISITIA 2017 - Proceeding*, *2017-Janua*, 356–361. https://doi.org/10.1109/ISITIA.2017.8124109

[10] Sahenda, L. N., Pumomo, M. H., Purnama, I. K. E., & Wisana, I. D. G. H. (2018). Comparison of Tuberculosis Bacteria Classification from Digital Image of Sputum Smears. *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia, CENIM 2018 - Proceeding*, 20–24. https://doi.org/10.1109/CENIM.2018.8711386

[11] Fitri, Z. E., Syahputri, L. N. Y., & Imron, A. M. N. (2020). Classification of White Blood Cell

Abnormalities for Early Detection of Myeloproliferative Neoplasms Syndrome Based on K-Nearest Neighborr. *Scientific Journal of Informatics*, *7*(1), 136–142. https://doi.org/10.15294/sji.v7i1.24372

[12] Fitri, Z. E., Sahenda, L. N., Puspitasari, P. S. D., Destarianto, P., Rukmi, D. L., & Imron, A. M. N. (2021). The Classification of Acute Respiratory Infection (ARI) Bacteria Based on K-Nearest Neighbor. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *12*(2), 91–101.

[13] Fitri, Z. E., Baskara, A., Silvia, M., Madjid, A., & Imron, A. M. N. (2021). Application of backpropagation method for quality sorting classification system on white dragon fruit (Hylocereus undatus). *IOP Conf. Series: Earth and Environmental Science*, *672*(IT Agriculture), 1–6. https://doi.org/10.1088/1755-1315/672/1/012085

[14] Fitri, Z. E., Nuhanatika, U., Madjid, A., & Imron, A. M. N. (2020). Penentuan Tingkat Kematangan Cabe Rawit (Capsicum frutescens L.) Berdasarkan Gray Level Co-Occurrence Matrix. *Jurnal Teknologi Informasi Dan Terapan*, *7*(1), 1–5. https://doi.org/10.25047/jtit.v7i1.121

[15] Fitri, Z. E., Rizkiyah, R., Madjid, A., & Imron, A. M. N. (2020). Penerapan Neural Network untuk Klasifkasi Kerusakan Mutu Tomat. *Jurnal Rekayasa Elektrika*, *16*(1), 44–49. https://doi.org/10.17529/jre.v16i1.15535

Prediction of the COVID-19 Vaccination Target Achievement with Exponential Regression

Teja Endra Eng Tju[1*], Dian Sa'adillah Maylawati[2], Ghifari Munawar[3], Suharjanto Utomo[4]
[1]Universitas Budi Luhur
[2]UIN Sunan Gunung Djati Bandung
[3]Politeknik Negeri Bandung
[4]Universitas Nurtanio
[*]Email: teja.endraengtju@budiluhur.ac.id

Abstract – The achievement of the national COVID-19 vaccination target in Indonesia is often reported to be uncertain with various existing obstacles. Prediction with exponential regression modeling is done by adopting part of the SKKNI Data Science with the stages of Data Understanding, Data Preparation, Modeling, Model Evaluation. The vaccination dataset from the Ministry of Health of the Republic of Indonesia for the period from January 13, 2021 to October 10, 2021, was randomly separated into training data of 0.8 parts and testing data of 0.2 parts. The optimal parameters of the exponential function are found using the scipy.optimize library in IPython. The model obtained was evaluated using MAE, RMSE, and R-Squared metrics on normalized training data, training data, test data, and recent data for seven days from 11 to 17 October 2021. The prediction results show that the vaccination target will be achieved 100 percent on January 18, 2022, while on December 31, 2021, only 80 percent will be achieved. From the recent data, it appears that more acceleration is needed, especially if it is desired to be achieved in December 2021 as determined by President Joko Widodo, there will be a shortfall of 20 percent based on the prediction results.

Keywords – Prediction, Exponential, Regression, Vaccination, COVID-19, SKKNI Data Science.

I. INTRODUCTION

The achievement of the target of the Indonesian population having been fully vaccinated of COVID-19 is not certain when it will occur. Some of these obstacles include resistance from the population, areas that are difficult to reach, inefficient vaccination implementation, distribution and availability of vaccines that are not smooth. The emergence of new virus variants makes it difficult to achieve the target of herd immunity due to vaccine efficacy constraints, so now the government has changed the vaccination target to a minimum of 208,265,720 Indonesians or about 80 percent of the total population so that the outbreak can be controlled [1]. Initially the target time was set for March 2022, then President Joko Widodo wanted an acceleration to December 2021 on the grounds that the economy could run. However, some officials and figures often mention different things according to the conditions at that time [2].

Previous research has mostly been done on the spread or increase of COVID-19 sufferers, not on vaccination. Prediction of the epidemic in Egypt is done by various regression analysis [3]. The spreading trend in China is used as an exponential attractor [4]. Modeling with reverse exponential regression for daily cases in Saudi Arabia [5]. In India, modeling and forecasting growth curves using various analytical techniques [6] as well as piecewise regression techniques [7]. For the prediction of COVID-19 cases in Indonesia using the hybrid method nonlinear regression logistic – double exponential smoothing [8], single exponential smoothing and the Holt's method [9], exponential smoothing method [10].

The Data Science approach that utilizes AI (Artificial Intelligence) technology can be a solution to present insight knowledge from data or facts related to vaccination rates in Indonesia so that it can be used as a basis or recommendation in decision making for policy makers. AI technology can help predict when Indonesia will be able to achieve herd immunity. This prediction can also lead to recommendations for further activities of the Indonesian people, such as policies related to health protocols, community activities in public places, activities in the work and school environment, and other policies related to recovering post-pandemic conditions or living side by side with Covid-19. These predictions can be processed with AI techniques in Data Science based on data on vaccination rates that have been carried out in various situations and conditions, so that from time to time it can be scientifically determined when the target number of vaccinations can be achieved. Likewise, if the target amount changes, the target time will be more easily analyzed. Therefore, this study aims to predict the achievement of the Covid-19 vaccination target in Indonesia.

II. RESEARCH METHODOLOGY

The dataset was obtained from the website of the Ministry of Health of the Republic of Indonesia which specifically provides vaccination report [12]. The data used for modeling is the second vaccination dose from January 13, 2021 to October 10, 2021 or as many as 271 data which is the result of grouping from various regions or population demographics in Indonesia.

The research methodology refers to the Indonesian National Work Competency Standard (SKKNI) No. 299 of 2020 in the field of Artificial Intelligence, sub-field of Data science [11]. The SKKNI Data Science consists of seven main activities, namely business understanding, data understanding, data preparation, modeling, model evaluation, deployment, and evaluation. This paper adopts four activities relevant to the research conducted, namely data understanding, data preparation, modeling, and model evaluation.

A. Data Understanding

The data is a time series with date as the independent variable and the second vaccination achievement as the dependent variable.

B. Data Preparation

For regression purposes, the date needs to be changed to an index starting from day 0 (zero) to day 270, sequentially according to the daily data date.

The dependent variable is the percentage of the daily cumulative amount, such that when it reaches 100 percent, it means that 208,265,720 or 80 percent of the total population of Indonesia have been vaccinated.

The dataset with 271 data was divided into two parts randomly, 0.8 part as the training dataset with 216 data and 0.2 part as the testing dataset with 55 data. The distribution of data and the results of data compilation are visualized [13] in Figure 1.

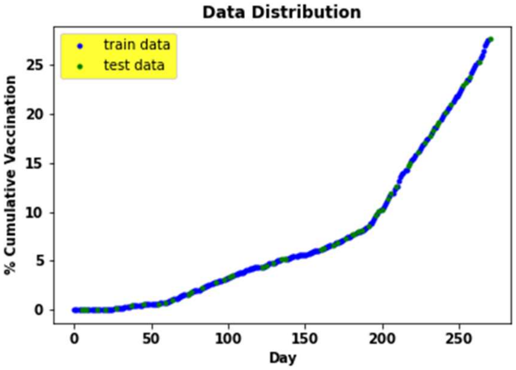


Figure 1. Data Distribution

C. Modeling

From the characteristics of the data with increasing growth from time to time, also driven by the government's desire to accelerate vaccination, the suitable modeling approach is Exponential Regression.

$$f(x) = ab^x + c \tag{1}$$

The Exponential Function [14] used is shown in Equation (1) where *x* is the independent variable, *f(x)* is the dependent variable, and there are three parameters *a, b, c* which the optimal values are sought with curve_fit from the scipy.optimize library [15] in IPython [16].

Data normalization was carried out [17] to obtain the appropriate model, after which the results were returned to the original values. The model will be built from normalized training dataset.

D. Model Evaluation

There are three metrics used for model evaluation, namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-Squared [18]. Evaluation is carried out on training dataset, testing dataset, and the recent dataset. Recent dataset is new data from October 11, 2021 to October 17, 2021.

Furthermore, it is necessary to find out when the 100 percent vaccination target is achieved by iterating to get the value of the independent variable.

III. Results And Discussion

The results of the model development with training data normalized to the exponential function Equation (1) resulted in the optimal fit parameter values *a* = 0.037263815344230705, *b* = 28.481521431056155, and *c* = -0.03569552348715502, the visualization results are in Figure 2. The result after returning to the original value is shown in Figure 3.

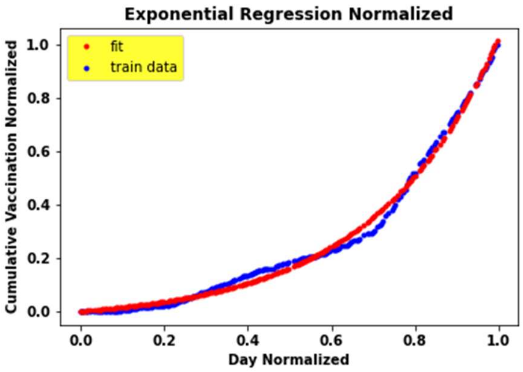


Figure 2. Exponential Regression Normalized

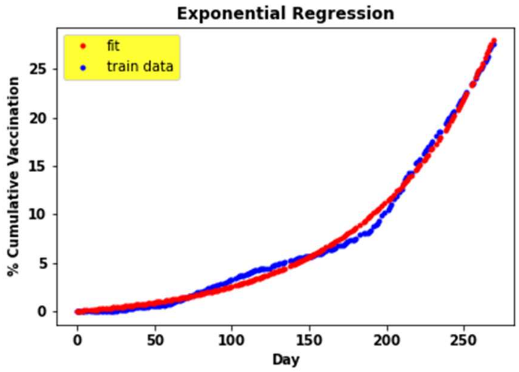


Figure 3. Exponential Regression

The results of the model evaluation on the testing dataset are visualized in Figure 4. Furthermore, the results on the new dataset for seven days from October 11, 2021 to October 17, 2021 can be seen in Figure 5.

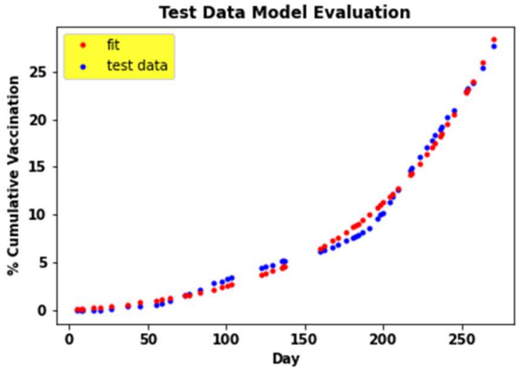


Figure 4. Test Data Model Evaluation

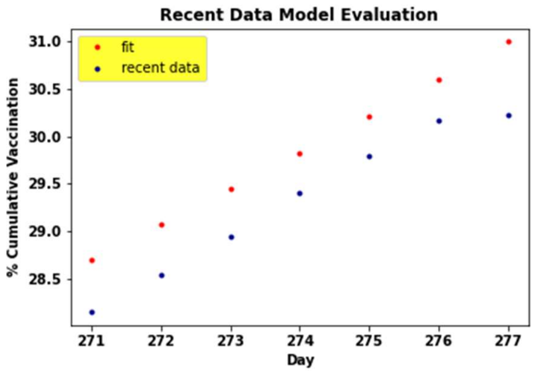


Figure 5. Recent Data Model Evaluation

Model Evaluation as measured by MAE, RMSE, R-Squared on training dataset, testing dataset, and the recent dataset is presented in Table 1. These results have been visualized in Figure 2, Figure 3, Figure 4, and Figure 5, respectively.

Table 1. Model Evaluations Result

Evaluation Metrics	Dataset			
	Normalized	Training	Testing	Recent
MAE	0.02	7.88	8.67	0.52
RMSE	0.02	10.62	10.97	0.53
R-Squared	0.99	0.99	0.99	0.51

To get the date when the vaccination target is reached 100 percent, an iteration has been carried out with the results on day 370 or on January 18, 2022. On the achievement of the recent data in Figure 5, it seems that it lags behind the fit data model (predicted results), so the implementation of daily vaccinations must be accelerated. Based on historical data, the acceleration of vaccination implementation should be maintained with reference to the predicted data, so it can be said as a prediction roadmap as shown in Figure 6. This prediction result is the optimal or moderate value of the government's target, which was originally March 2022 and then advanced to December 2021. On December 31, 2021 or day 352, 80 percent vaccination progress will be made, the Indonesian government needs to make efforts to accelerate in such a way that recent data can more often exceed the predicted results so that the December 2021 target can be achieved.

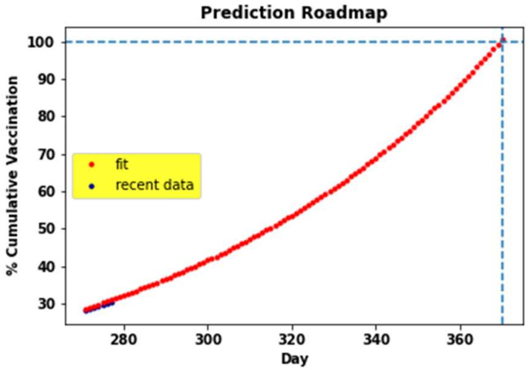


Figure 6. Prediction Roadmap

IV. Conclusion

Regression analytic approach based on historical data can be used to build a model for prediction. The characteristics of the data become an important asset, especially in terms of data growth to determine a suitable function as a model.

With exponential regression modeling able to predict the target of achieving COVID-19 vaccination in Indonesia. Functional regression can be used on continuous time series data in a variety of other problems to predict future results.

Based on predictions using the exponential regression model, the vaccination target will be reached 100 percent on January 18, 2022, while the government target for December 2021 has only reached 80 percent. This has an impact that the government needs to make various acceleration efforts so that herd immunity can occur in December 2021.

References

[1] Satuan Tugas Penanganan COVID-19, “208.265.720 Orang Target Sasaran Vaksinasi COVID-19 di Indonesia,” 2021. https://covid19.go.id/p/vaksin/208265720-orang-target-sasaran-vaksinasi-covid-19-di-indonesia.

[2] Kompas, “Sulit Kejar Target Vaksinasi Covid-19, Menkes: Kami Butuh Bantuan Pemda dan Swasta,” 2021. https://megapolitan.kompas.com/read/2021/06/02/13554891/sulit-kejar-target-vaksinasi-covid-19-menkes-kami-butuh-bantuan-pemda-dan-swasta.

[3] L. A. Amar, A. A. Taha, and M. Y. Mohamed, “Prediction of the final size for COVID-19 epidemic using machine learning: A case study of Egypt,” *Infect. Dis. Model.*, vol. 5, 2020, doi: 10.1016/j.idm.2020.08.008.

[4] C. C. Zhu and J. Zhu, “Spread trend of COVID-19

epidemic outbreak in China: Using exponential attractor method in a spatial heterogeneous SEIQR model," *Math. Biosci. Eng.*, vol. 17, no. 4, 2020, doi: 10.3934/MBE.2020174.

[5] S. R. Al-Dawsari and K. S. Sultan, "Modeling of daily confirmed Saudi COVID-19 cases using inverted exponential regression," *Math. Biosci. Eng.*, vol. 18, no. 3, 2021, doi: 10.3934/MBE.2021117.

[6] V. K. Sharma and U. Nigam, "Modeling and Forecasting of COVID-19 Growth Curve in India," *Trans. Indian Natl. Acad. Eng.*, vol. 5, no. 4, 2020, doi: 10.1007/s41403-020-00165-z.

[7] A. Senapati, A. Nag, A. Mondal, and S. Maji, "A novel framework for COVID-19 case prediction through piecewise regression in India," *Int. J. Inf. Technol.*, vol. 13, no. 1, 2021, doi: 10.1007/s41870-020-00552-3.

[8] I. G. B. Ngurah Diksa, "Peramalan Gelombang Covid 19 Menggunakan Hybrid Nonlinear Regression Logistic – Double Exponential Smoothing di Indonesia dan Prancis," *Jambura J. Math.*, vol. 3, no. 1, 2021, doi: 10.34312/jjom.v3i1.7771.

[9] N. H. A. S. Al Ihsan, H. H. Dzakiyah, and F. Liantoni, "Perbandingan Metode Single Exponential Smoothing dan Metode Holt untuk Prediksi Kasus COVID-19 di Indonesia," *Ultim. J. Tek. Inform.*, vol. 12, no. 2, 2020, doi: 10.31937/ti.v12i2.1689.

[10] C. M. Gibran, S. Setiyawati, and F. Liantoni, "Prediksi Penambahan Kasus Covid-19 di Indonesia Melalui Pendekatan Time Series Menggunakan Metode Exponential Smoothing," *J. Inform. Univ. Pamulang*, vol. 6, no. 1, 2021, doi: 10.32493/informatika.v6i1.9442.

[11] Kementrian Ketenagakerjaan Republik Indonesia, "SKKNI Keahlian Artificial Intelligence (Data Science)," 2020. https://skkni.kemnaker.go.id/tentang-skkni/dokumen?area=data science&limit=20&page=1.

[12] Kementrian Kesehatan Republik Indonesia, "Vaksinansi COVID-19 Nasional," 2021. https://vaksin.kemkes.go.id/#/vaccines.

[13] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[14] R. Torres, "Mathematical Investigation of Functions," *J. Educ. Manag. Dev. Stud.*, vol. 1, no. 1, 2021, doi: 10.52631/jemds.v1i1.6.

[15] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nat. Methods*, vol. 17, no. 3, pp. 261–272, 2020, doi: 10.1038/s41592-019-0686-2.

[16] P. Fernando and E. G. Brian, "IPython: A System for Interactive Scientific Computing," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 21–29, 2007.

[17] S. Lakshmanan, "How, When, and Why Should You Normalize / Standardize / Rescale Your Data?," *Towards AI — The Best of Tech, Science, and Engineering*, 2019. https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff.

[18] J. Hale, "Which Evaluation Metric Should You Use in Machine Learning Regression Problems?," *Towords Datascience*, 2020. https://towardsdatascience.com/which-evaluation-metric-should-you-use-in-machine-learning-regression-problems-20cdaef258e.